Advanced High School Statistics

Third Edition in Progress,

NOT FOR DISTRIBUTION

David Diez
Data Scientist
OpenIntro

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

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

Christopher D Barr
Investment Analyst
Varadero Capital

Updated: August 20th, 2019.

Table of Contents

2 Summarizing data **6**
2.1 Examining numerical data . . . 8
2.1.1 Scatterplots for paired data . . . 9
2.1.2 Stem-and-leaf plots and dot plots . . . 11
2.1.3 Histograms . . . 13
2.1.4 Describing Shape . . . 16
2.1.5 Descriptive versus inferential statistics . . . 17
2.2 Numerical summaries and box plots . . . 22
2.2.1 Measures of center . . . 22
2.2.2 Standard deviation as a measure of spread . . . 25
2.2.3 Z-scores . . . 28
2.2.4 Box plots and quartiles . . . 29
2.2.5 Technology: summarizing 1-variable statistics . . . 32
2.2.6 Outliers and robust statistics . . . 35
2.2.7 Linear transformations of data . . . 37
2.2.8 Comparing numerical data across groups . . . 38
2.2.9 Mapping data (special topic) . . . 41
2.3 Normal distribution . . . 50
2.3.1 Normal distribution model . . . 50
2.3.2 Using the normal distribution to approximate empirical distributions . . . 52
2.3.3 Finding areas under the normal curve . . . 53
2.3.4 Normal probability examples . . . 54
2.3.5 Evaluating the normal approximation (special topic) . . . 57
2.3.6 Technology: finding normal probabilities . . . 58
2.4 Considering categorical data . . . 64
2.4.1 Contingency tables and bar charts . . . 64
2.4.2 Row and column proportions . . . 66
2.4.3 Using a bar chart with two variables . . . 68
2.4.4 Mosaic plots . . . 69
2.4.5 The only pie chart you will see in this book . . . 71
2.5 Case study: malaria vaccine (special topic) . . . 74
2.5.1 Variability within data . . . 74
2.5.2 Simulating the study . . . 76
2.5.3 Checking for independence . . . 76
Chapter highlights . . . 80
Chapter exercises . . . 81

3 Probability and probability distributions **83**
3.1 Defining probability . . . 85
3.1.1 Introductory examples . . . 85
3.1.2 Probability . . . 86
3.1.3 Disjoint or mutually exclusive outcomes . . . 87
3.1.4 Probabilities when events are not disjoint . . . 89

3.1.5 Complement of an event . . . 91
3.1.6 Independence . . . 92
3.2 Conditional probability . . . 99
3.2.1 Exploring probabilities with a contingency table . . . 100
3.2.2 Marginal and joint probabilities . . . 101
3.2.3 Defining conditional probability . . . 102
3.2.4 Smallpox in Boston, 1721 . . . 104
3.2.5 General multiplication rule . . . 105
3.2.6 Sampling without replacement . . . 106
3.2.7 Independence considerations in conditional probability . . . 108
3.2.8 Checking for independent and mutually exclusive events . . . 108
3.2.9 Tree diagrams . . . 111
3.2.10 Bayes' Theorem . . . 112
3.3 Simulations . . . 121
3.3.1 Setting up and carrying out simulations . . . 121
3.4 Random variables . . . 127
3.4.1 Introduction to expected value . . . 127
3.4.2 Probability distributions . . . 128
3.4.3 Expectation . . . 130
3.4.4 Variability in random variables . . . 132
3.4.5 Linear transformations of a random variable . . . 133
3.4.6 Linear combinations of random variables . . . 134
3.4.7 Variability in linear combinations of random variables . . . 136
3.4.8 Normal approximation for sums of random variables . . . 138
3.5 Geometric distribution . . . 143
3.5.1 Bernoulli distribution . . . 143
3.5.2 Geometric distribution . . . 144
3.5.3 Technology: geometric probablities . . . 146
3.6 Binomial distribution . . . 149
3.6.1 Introducing the binomial formula . . . 149
3.6.2 When and how to apply the formula . . . 151
3.6.3 Technology: binomial probabilities . . . 154
3.6.4 An example of a binomial distribution . . . 156
3.6.5 The mean and standard deviation of a binomial distribution . . . 156
3.6.6 Normal approximation to the binomial distribution . . . 157
3.6.7 Normal approximation breaks down on small intervals (special topic) . . . 160
Chapter highlights . . . 165
Chapter exercises . . . 166

8 Introduction to linear regression 168
8.1 Line fitting, residuals, and correlation . . . 170
8.1.1 Fitting a line to data . . . 170
8.1.2 Using linear regression to predict possum head lengths . . . 172
8.1.3 Residuals . . . 174
8.1.4 Describing linear relationships with correlation . . . 178
8.2 Fitting a line by least squares regression . . . 187
8.2.1 An objective measure for finding the best line . . . 187
8.2.2 Finding the least squares line . . . 189
8.2.3 Interpreting the coefficients of a regression line . . . 191
8.2.4 Extrapolation is treacherous . . . 192
8.2.5 Using R^2 to describe the strength of a fit . . . 193
8.2.6 Technology: linear correlation and regression . . . 195
8.2.7 Types of outliers in linear regression . . . 198
8.2.8 Categorical predictors with two levels (special topic) . . . 200
8.3 Transformations for skewed data . . . 206
8.3.1 Introduction to transformations . . . 206
8.3.2 Transformations to achieve linearity . . . 208
8.4 Inference for the slope of a regression line . . . 213

8.4.1 The role of inference for regression parameters . . . 213
8.4.2 Conditions for the least squares line . . . 214
8.4.3 Constructing a confidence interval for the slope of a regression line . . . 215
8.4.4 Technology: the t-interval for the slope . . . 219
8.4.5 Midterm elections and unemployment . . . 220
8.4.6 Understanding regression output from software . . . 222
8.4.7 Technology: the t-test for the slope . . . 226
8.4.8 Which inference procedure to use for paired data? . . . 227
Chapter highlights . . . 233
Chapter exercises . . . 234

A Exercise solutions **237**

B Data sets within the text **258**

C Distribution tables **263**

D Calculator reference, Formulas, and Inference guide **274**

Preface

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

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

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

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

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

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

Textbook overview

The chapters of this book are as follows:

1. **Data collection.** Data structures, variables, and basic data collection techniques.
2. **Summarizing data.** Data summaries and graphics.
3. **Probability.** The basic principles of probability.
4. **Distributions of random variables.** Introduction to key distributions, and how the normal model applies to the sample mean and sample proportion.
5. **Foundations for inference.** General ideas for statistical inference in the context of estimating the population proportion.
6. **Inference for categorical data.** Inference for proportions and contingency tables using the normal and chi-square distributions.
7. **Inference for numerical data.** Inference for one or two sample means using the t-distribution.
8. **Introduction to linear regression.** An introduction to regression with two variables, and inference on the slope of the regression line.

Online resources

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

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

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

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

All of these resources can be found at:

openintro.org/ahss

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

Examples and exercises

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

EXAMPLE 0.1

(E) This is an example.

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

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

GUIDED PRACTICE 0.2

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

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

Getting involved

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

openintro.org/ahss/feedback

Acknowledgements

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

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

Chapter 2

Summarizing data

2.1 Examining numerical data

2.2 Numerical summaries and box plots

2.3 Normal distribution

2.4 Considering categorical data

2.5 Case study: malaria vaccine (special topic)

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

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

2.1 Examining numerical data

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

Learning objectives

1. Use scatterplots to represent bivariate data and to see the relationship between two numerical variables. Describe the direction, form, and strength of the relationship, as well as any unusual observations.
2. Understand what the term distribution means and how to summarize it in a table or a graph.
3. Create univariate displays, including stem-and-leaf plots, dot plots, and histograms, to visualize the distribution of a numerical variable. Be able to read off specific information and summary information from these graphs.
4. Identify the shape of a distribution as approximately symmetric, right skewed, or left skewed. Also, identify whether a distribution is unimodal, bimodal, multimodal, or uniform.
5. Read and interpret a cumulative frequency or cumulative relative frequency histogram.

2.1.1 Scatterplots for paired data

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

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

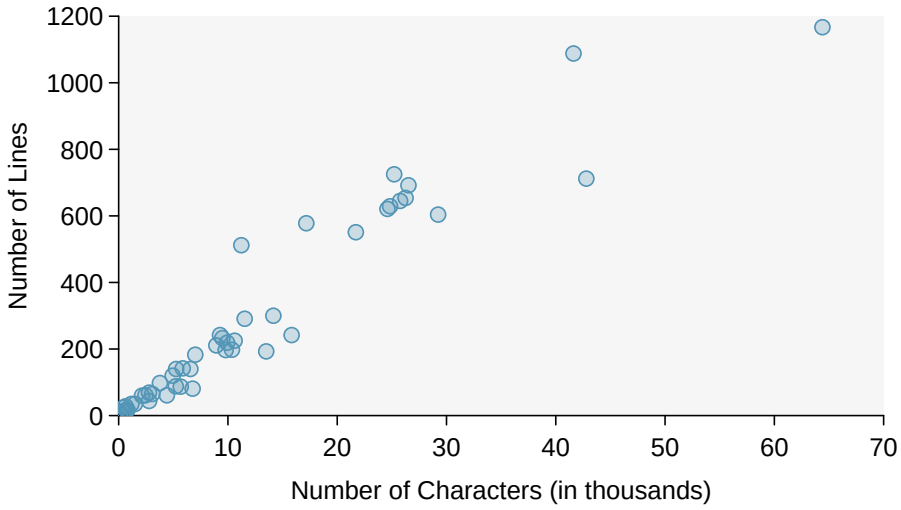


Figure 2.1: A scatterplot of `line_breaks` versus `num_char` for the `email50` data.

EXAMPLE 2.1

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

(E)

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

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

DRAWING SCATTERPLOTS

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

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

EXAMPLE 2.2

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

(E)

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

EXAMPLE 2.3

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

(E)

The relationship is evidently **nonlinear**, as highlighted by the dashed line. This is different from previous scatterplots we've seen, which show relationships that do not show much, if any, curvature in the trend. There is also a negative association, as higher rates of poverty tend to be associated with lower median household income.

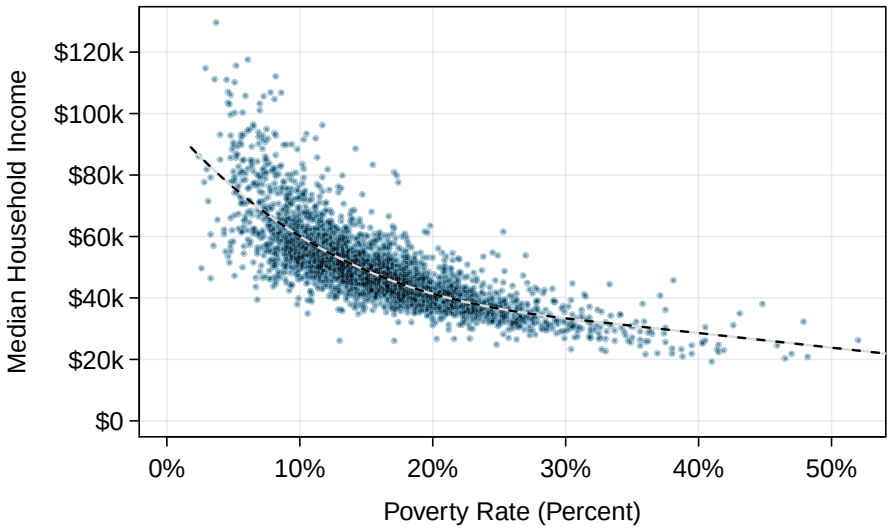


Figure 2.2: A scatterplot of the median household income against the poverty rate for the `county` data set. A statistical model has also been fit to the data and is shown as a dashed line. Explore dozens of scatterplots using American Community Survey data on Tableau Public.

(G)

GUIDED PRACTICE 2.4

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

(G)

GUIDED PRACTICE 2.5

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

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

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

2.1.2 Stem-and-leaf plots and dot plots

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

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

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

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

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

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

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

G **GUIDED PRACTICE 2.6**

There are a lot of numbers on the first row of the stem-and-leaf plot. Why is this the case?[3]

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

G **GUIDED PRACTICE 2.7**

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

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

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

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

Figure 2.5: A split stem-and-leaf.

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

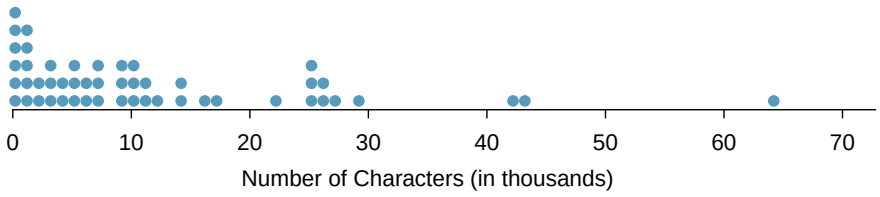


Figure 2.6: A dot plot of `num_char` for the `email50` data set.

G

GUIDED PRACTICE 2.8

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

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

EXAMPLE 2.9

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

E

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

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

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

OUTLIERS ARE EXTREME

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

WHY IT IS IMPORTANT TO LOOK FOR OUTLIERS

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

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

GUIDED PRACTICE 2.10

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

GUIDED PRACTICE 2.11

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

2.1.3 Histograms

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

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

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

GUIDED PRACTICE 2.12

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

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

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

[8] Character counts for individual emails.

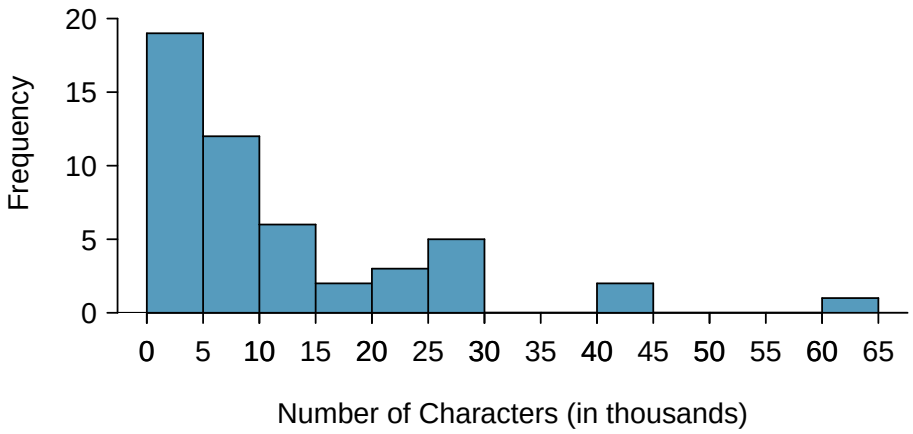


Figure 2.8: A histogram of `num_char`. This histogram uses bins or class intervals of width 5. Explore this histogram and dozens of histograms using American Community Survey data on Tableau Public.

DRAWING HISTOGRAMS

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

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

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

EXAMPLE 2.13

How many emails had fewer than 10 thousand characters?

(E)

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

EXAMPLE 2.14

Approximately how many emails had fewer than 1 thousand chacters?

(E)

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

EXAMPLE 2.15

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

(E)

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

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

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

Figure 2.9: The cumulative frequencies for the binned `num_char` data.

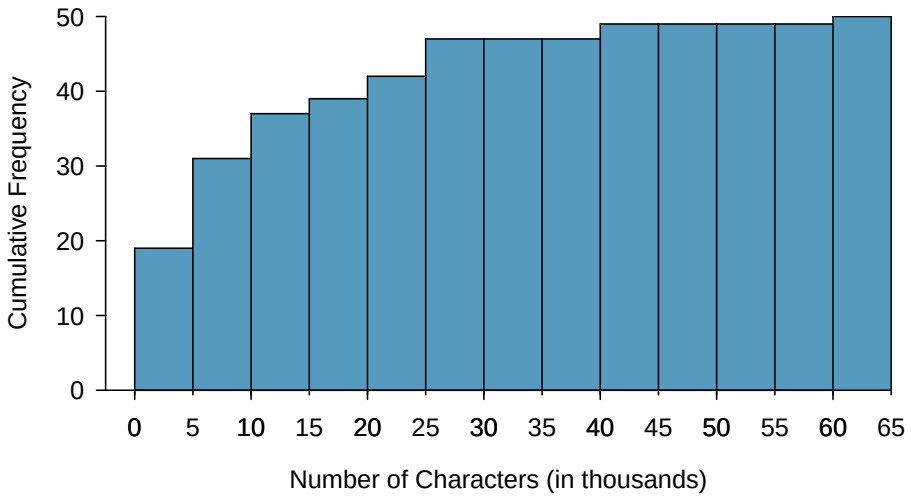


Figure 2.10: A cumulative frequency histogram of `num_char`. This histogram uses bins or class intervals of width 5. Compare frequency, relative frequency, cumulative frequency, and cumulative relative frequency histograms on Tableau Public.

EXAMPLE 2.16

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

(E)

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

EXAMPLE 2.17

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

(E)

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

EXAMPLE 2.18

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

(E)

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

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

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

GUIDED PRACTICE 2.19

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

PAY CLOSE ATTENTION TO THE VERTICAL AXIS OF A HISTOGRAM

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

2.1.4 Describing Shape

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

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

LONG TAILS TO IDENTIFY SKEW

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

GUIDED PRACTICE 2.20

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

GUIDED PRACTICE 2.21

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

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

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

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

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

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

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

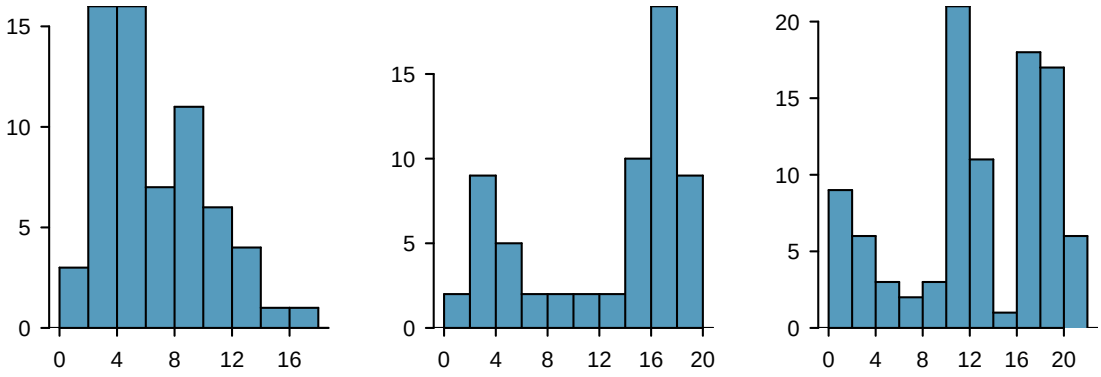


Figure 2.11: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

GUIDED PRACTICE 2.22

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

LOOKING FOR MODES

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

2.1.5 Descriptive versus inferential statistics

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

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

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

Section summary

- A **scatterplot** is a **bivariate** display illustrating the relationship between two numerical variables. The observations must be **paired**, which is to say that they correspond to the same case or individual. The linear association between two variables can be positive or negative, or there can be no association. **Positive association**positive association means that larger values of the first variable are associated with larger values of the second variable. **Negative association**negative association means that larger values of the first variable are associated with smaller values of the second variable. Additionally, the association can follow a linear trend or a curved (nonlinear) trend.
- When looking at a **univariate** display, researchers want to understand the distribution of the variable. The term **distribution** refers to the values that a variable takes and the frequency of those values. When looking at a distribution, note the presence of clusters, gaps, and **outliers**outlier.
- Distributions may be **symmetric** or they may have a long tail. If a distribution has a long left tail (with greater density over the higher numbers), it is **left skewed**. If a distribution has a long right tail (with greater density over the smaller numbers), it is **right skewed**.
- Distributions may be **unimodal**, **bimodal**, or **multimodal**.
- Two graphs that are useful for showing the distribution of a small number of observations are the **stem-and-leaf plot** and **dot plot**. These graphs are ideal for displaying data from small samples because they show the exact values of the observations and how frequently they occur. However, they are impractical for larger data sets.
- For larger data sets it is common to use a **frequency histogram** or a **relative frequency histogram** to display the distribution of a variable. This requires choosing bins of an appropriate width.
- To see cumulative amounts, use a **cumulative frequency histogram**. A **cumulative relative frequency histogram** is ideal for showing **percentile**.
- **Descriptive statistics** describes or summarizes data, while **inferential statistics** uses samples to generalize or infer something about a larger population.

Exercises

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

	Income	Age	Gender
1	53,000	28	male
2	1600	18	female
3	70,000	54	male
4	12,800	22	male
5	1,200	18	female
6	30,000	34	male
7	4,500	21	male
8	20,000	28	female
9	25,000	29	female
10	42,000	33	male

	Income	Age	Gender
11	670	34	female
12	29,000	55	female
13	44,000	33	female
14	48,000	41	male
15	30,000	47	female
16	60,000	30	male
17	108,000	61	male
18	5,800	50	female
19	50,000	24	female
20	11,000	19	male

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

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

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

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

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

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

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

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

[15] **data:acs:2012**.

[16] **data:MLB:2014**.

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

	Brand	Fiber %
1	Pebbles Fruity	0.0%
2	Rice Krispies Treats	0.0%
3	Pebbles Cocoa	0.0%
4	Pebbles Marshmallow	0.0%
5	Frosted Rice Krispies	0.0%
6	Rice Krispies	3.0%
7	Trix	3.1%
8	Honey Comb	3.1%
9	Rice Krispies Gluten Free	3.3%
10	Frosted Flakes	3.3%

	Brand	Fiber %
11	Cinnamon Toast Crunch	3.3%
12	Reese's Puffs	3.4%
13	Cheerios Honey Nut	7.1%
14	Lucky Charms	7.4%
15	Pebbles Boulders Chocolate PB	7.4%
16	Corn Pops	9.4%
17	Frosted Flakes Reduced Sugar	10.0%
18	Clifford Crunch	10.0%
19	Apple Jacks	10.7%
20	Dora the Explorer	11.1%

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

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

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

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

2.4 Sugar in your cereal. The Cereal FACTS report from Exercise 2.3 also provides information on sugar content of cereals. We have selected a random sample of 20 cereals from the data provided in this report. Shown below are the sugar contents (percentage of sugar per gram of cereal) for these cereals.

	Brand	Sugar %
1	Rice Krispies Gluten Free	3%
2	Rice Krispies	12%
3	Dora the Explorer	22%
4	Frosted Flakes Red. Sugar	27%
5	Clifford Crunch	27%
6	Rice Krispies Treats	30%
7	Pebbles Boulders Choc. PB	30%
8	Cinnamon Toast Crunch	30%
9	Trix	31%
10	Honey Comb	31%

	Brand	Sugar %
11	Corn Pops	31%
12	Cheerios Honey Nut	32%
13	Reese's Puffs	34%
14	Pebbles Fruity	37%
15	Pebbles Cocoa	37%
16	Lucky Charms	37%
17	Frosted Flakes	37%
18	Pebbles Marshmallow	37%
19	Frosted Rice Krispies	40%
20	Apple Jacks	43%

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

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

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

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

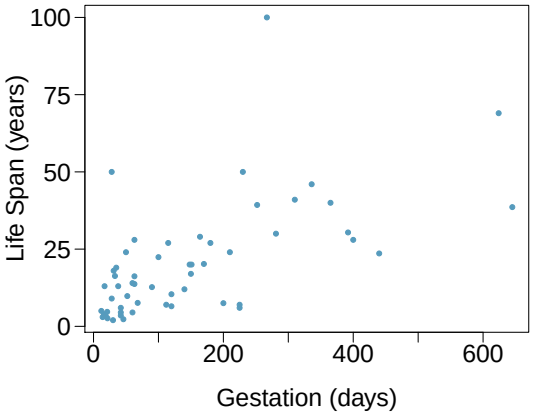
[17] **Harris:2012**.

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

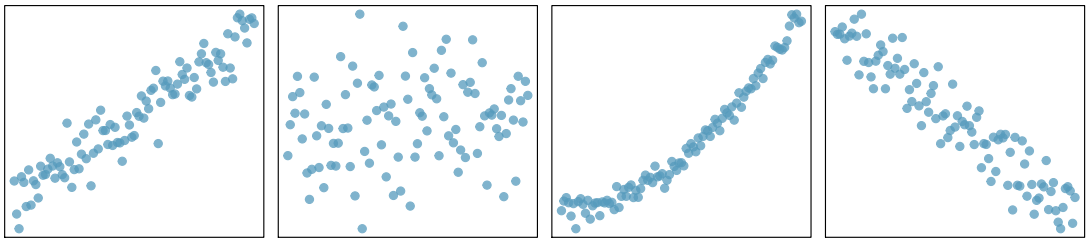
(a) What type of an association is apparent between life span and length of gestation?

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

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



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



[18] **Allison+Cicchetti:1975**.

2.2 Numerical summaries and box plots

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

Learning objectives

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

2.2.1 Measures of center

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

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

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

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

MEAN

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

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

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

GUIDED PRACTICE 2.23

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

GUIDED PRACTICE 2.24

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

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

EXAMPLE 2.25

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

E

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 **??** and beyond, we will develop tools to characterize the reliability of point estimates, and we will find that point estimates based on larger samples tend to be more reliable than those based on smaller samples.

EXAMPLE 2.26

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

E

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

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

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

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

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

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

MEDIAN: THE NUMBER IN THE MIDDLE

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

Graphically, we can think of the mean as the balancing point. The median is the value such that 50% of the *area* is to the left of it and 50% of the *area* is to the right of it.

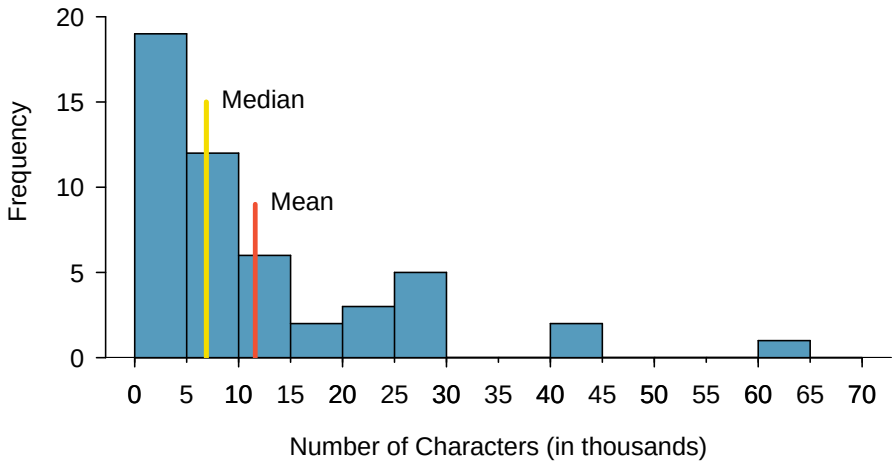


Figure 2.12: A histogram of `num_char` with its mean and median shown.

EXAMPLE 2.27

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

(E)

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

THE MEAN FOLLOWS THE TAIL

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

GUIDED PRACTICE 2.28

(G)

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

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

2.2.2 Standard deviation as a measure of spread

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

EXAMPLE 2.29

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

(E)

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

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

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

We call the distance of an observation from its mean its **deviation**. Below are the deviations for the 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:

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

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:

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

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

[22] https://data.census.gov/cedsci/table?hidePreview=true&tid=ACSST1Y2017.S1901

CALCULATING THE STANDARD DEVIATION

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

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

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

THINKING ABOUT THE STANDARD DEVIATION

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

EXAMPLE 2.30

Based on American Community Survey data, the mean family income in the U.S. in 2017 was \$99,114.[24] Estimating the standard deviation of income as approximately \$50,000, is a family income of \$60,000 far from the mean or relatively close to the mean?

(E)

Because \$60,000 is less that one standard deviation from the mean, it is relatively close to the mean. If the value were more than 2 standard deviations away from the mean, we would consider it far from the mean.

In the next section, we encounter a bell-shaped distribution known as the *normal distribution*. The **empirical rule** tells us that for nearly normal distributions, about 68% of the data will be within one standard deviation of the mean, about 95% will be within two standard deviations of the mean, and about 99.7% will be within three standard deviations of the mean. However, as seen in Figures 2.13 and 2.14, these percentages generally do not hold if the distribution is not bell-shaped.

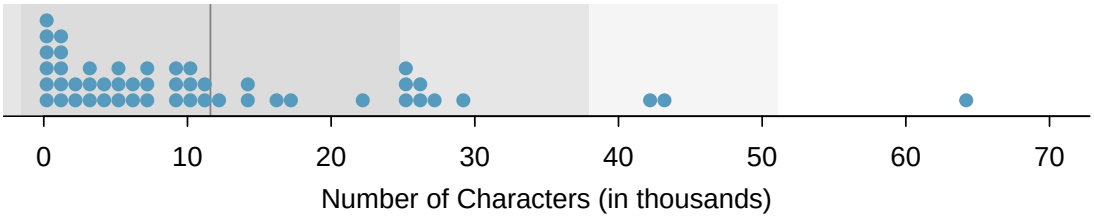


Figure 2.13: In the `num_char` data, 40 of the 50 emails (80%) are within 1 standard deviation of the mean, and 47 of the 50 emails (9 4%) are within 2 standard deviations. The empirical rule does not hold well for skewed data, as shown in this example.

GUIDED PRACTICE 2.31

(G)

On page 16, the concept of shape of a distribution was introduced. A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using Figure 2.14 as an example, explain why such a description is important.[25]

[23] The only difference is that the population variance has a division by n instead of $n-1$.
[24] https://data.census.gov/cedsci/table?hidePreview=true&tid=ACSST1Y2017.S1901
[25] Figure 2.14 shows three distributions that look quite different, but all have the same mean, variance, and standard

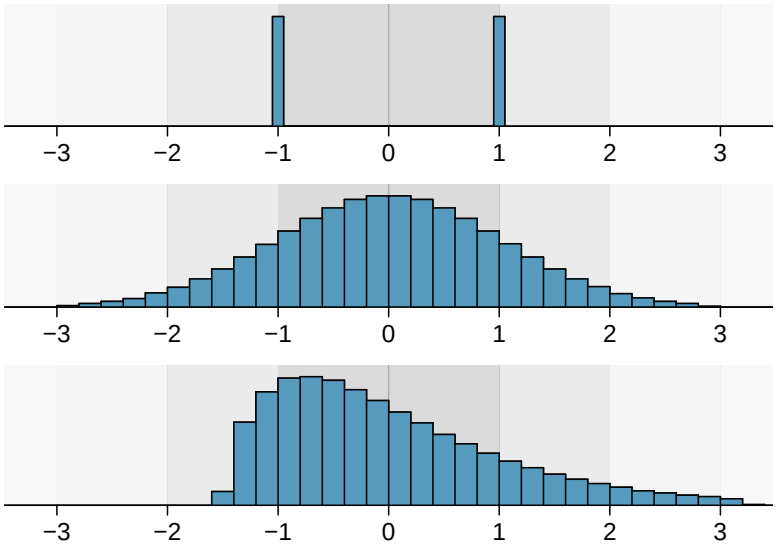
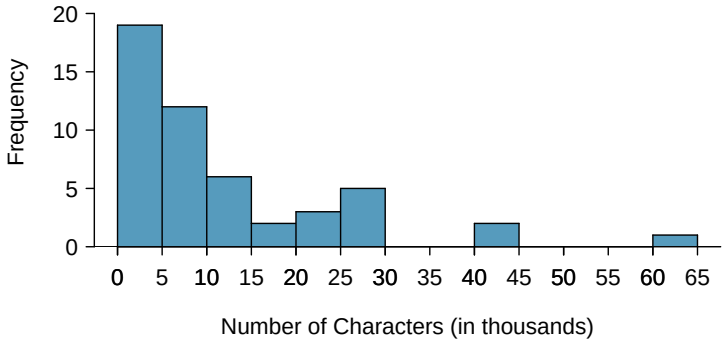


Figure 2.14: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

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

EXAMPLE 2.32

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



E

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

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

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.

2.2.3 Z-scores

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

EXAMPLE 2.33

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

(E)

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

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

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

THE Z-SCORE

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

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

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

EXAMPLE 2.34

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

(E)

For $x_1 = 95.4$ mm:

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

For $x_2 = 85.8$ mm:

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

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

G

GUIDED PRACTICE 2.35

Which of the observations in Example 2.34 is more unusual?[26]

G

GUIDED PRACTICE 2.36

Let X represent a random variable from a distribution with $\mu = 3$ and $\sigma = 2$, and suppose we observe $x = 5.19$.
(a) Find the Z-score of x.
(b) Interpret the Z-score.[27]

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

EXAMPLE 2.37

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

E

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

2.2.4 Box plots and quartiles

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

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

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

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

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

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

[27] (a) Its Z-score is given by $Z = \frac{x-\mu}{\sigma} = \frac{5.19-3}{2} = 2.19/2 = 1.095$. (b) The observation x is 1.095 standard deviations *above* the mean. We know it must be above the mean since Z is positive.

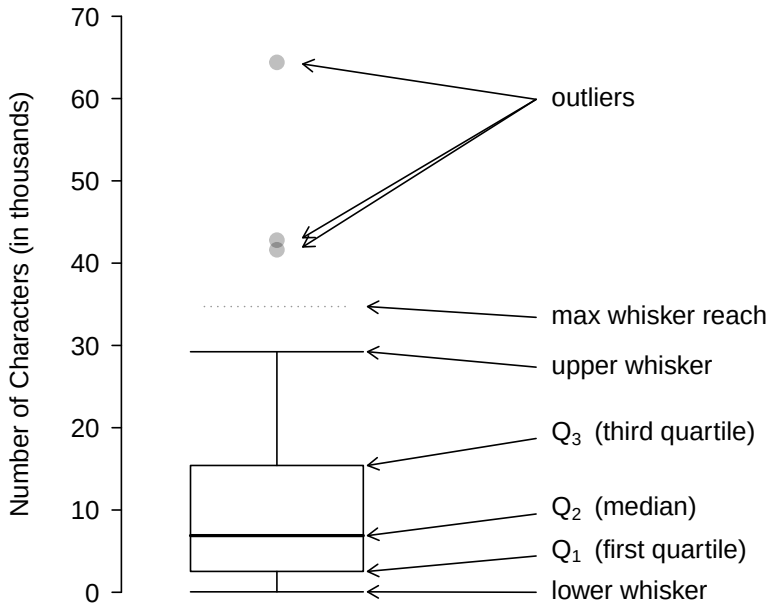


Figure 2.15: A labeled box plot for the number of characters in 50 emails. The median (6,890) splits the data into the bottom 50% and the top 50%. Explore dozens of boxplots with histograms using American Community Survey data on Tableau Public.

INTERQUARTILE RANGE (IQR)

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

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 are the 25^{th} and 75^{th} percentiles.

OUTLIERS IN THE CONTEXT OF A BOX PLOT

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

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

Extending out from the rectangle, the **whiskers** attempt to capture all of the data remaining outside of the box, except outliers. In Figure 2.15, the upper whisker does not extend to the last three points, which are beyond $Q_3 + 1.5 \times IQR$ and are outliers, so it extends only to the last point below this limit.[28] The lower whisker stops at the lowest value, 33, since there are no additional data to reach. Outliers are each marked with a dot or asterisk. In a sense, the box is like the body of the box plot and the whiskers are like its arms trying to reach the rest of the data.

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

EXAMPLE 2.38

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

(E)

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

EXAMPLE 2.39

Is it possible to identify skew from the box plot?

(E)

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

(G)

GUIDED PRACTICE 2.40

True or false: there is more data between the median and Q_3 than between Q_1 and the median.[29]

EXAMPLE 2.41

Consider the following ordered data set.

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

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

(E)

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

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

On the high end we need:

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

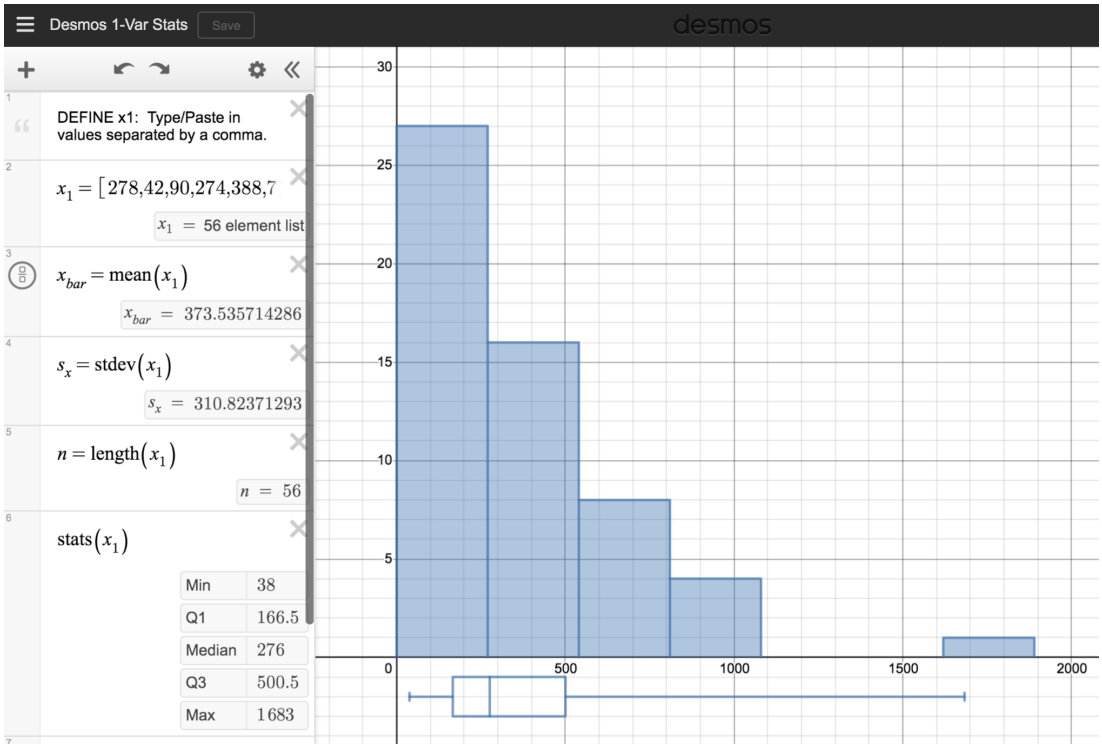
There are no numbers less than -41, so there are no outliers on the low end. The observation at 80 is greater than 77, so 80 is an outlier on the high end.

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

2.2.5 Technology: summarizing 1-variable statistics

Online calculators such as Desmos or a handheld calculator can be used to calculate summary statistics. More advanced statistical software packages include R (which was used for most of the graphs in this text), Python, SAS, and STATA.

Get started quickly with this Desmos 1-VarStats Calculator.[30]



Calculator instructions

TI-83/84: ENTERING DATA

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

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

CASIO FX-9750GII: ENTERING DATA

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

[30] Find this other Desmos Calculators and Activities referenced in the textbook at openintro.org/ahss/desmos.

TI-84: CALCULATING SUMMARY STATISTICS

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

1. Enter the data as described previously.
2. Press STAT.
3. Right arrow to CALC.
4. Choose 1:1-Var Stats.
5. Enter L1 (i.e. 2ND 1) for List. If the data is in a list other than L1, type the name of that list.
6. Leave FreqList blank.
7. Choose Calculate and hit ENTER.

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

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

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

TI-83/84: DRAWING A BOX PLOT

1. Enter the data to be graphed as described previously.
2. Hit 2ND Y= (i.e. STAT PLOT).
3. Hit ENTER (to choose the first plot).
4. Hit ENTER to choose ON.
5. Down arrow and then right arrow three times to select box plot with outliers.
6. Down arrow again and make Xlist: L1 and Freq: 1.
7. Choose ZOOM and then 9:ZoomStat to get a good viewing window.

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

Restore lists L1-L6 using the following steps:

1. Press STAT.
2. Choose 5:SetUpEditor.
3. Hit ENTER.

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

1. Navigate to STAT (MENU, then hit 2) and enter the data into a list.
2. Go to GRPH (F1).
3. Next go to SET (F6) to set the graphing parameters.
4. To use the 2nd or 3rd graph instead of GPH1, select F2 or F3.
5. Move down to Graph Type and select the ▷ (F6) option to see more graphing options, then select Box (F2).
6. If XList does not show the list where you entered the data, hit LIST (F1) and enter the correct list number.
7. Leave Frequency at 1.
8. For Outliers, choose On (F1).
9. Hit EXE and then choose the graph where you set the parameters F1 (most common), F2, or F3.
10. If desired, explore 1-variable statistics by selecting 1-Var (F1).

CASIO FX-9750GII: DELETING A DATA LIST

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

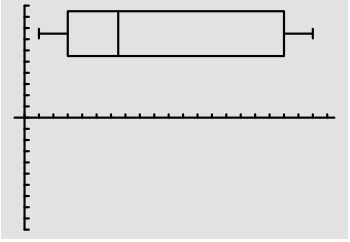
GUIDED PRACTICE 2.42

Enter the following 10 data points into a calculator.

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

Find the summary statistics and make a box plot of the data.[31]

[31] The summary statistics should be $\bar{x} = 9.1$, Sx $= 7.48$, Q1 $= 3$, etc. Using a TI, the boxplot looks like this:



GUIDED PRACTICE 2.43

Use the `email50` data set at openintro.org/data and Screen 2 of this Desmos 1-Var Stats Calculator to summarize the num_char variable (number of characters in an email).[32]

2.2.6 Outliers and robust statistics

RULES OF THUMB FOR IDENTIFYING OUTLIERS

There are two rules of thumb for identifying outliers:

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

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

GUIDED PRACTICE 2.44

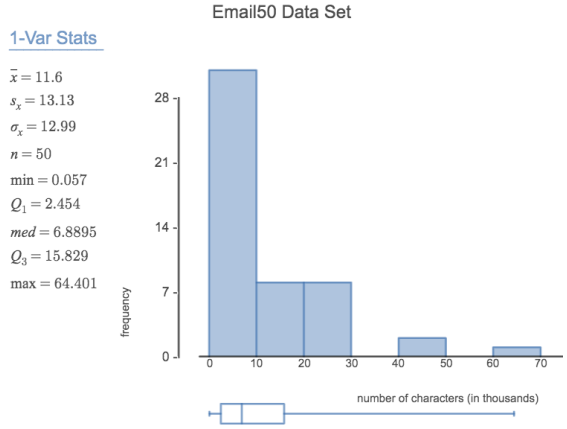
For the `email50` data set,$Q_1 = 2{,}536$ and $Q_3 = 15,411$. $\bar{x} = 11{,}600$ and $s = 13{,}130$. What values would be considered an outlier on the low end using each rule?[33]

GUIDED PRACTICE 2.45

Because there are no negative values in this data set, there can be no outliers on the low end. What does the fact that there are outliers on the high end but not on the low end suggestion?[34]

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 2.16, and sample statistics are computed under each scenario in Figure 2.17.

[32]Remember, the Desmos Calculators and Activities in this book can be found at openintro.org/ahss/desmos. Down the `email50` CSV file and open it. Copy and paste the num_char column into the Desmos calculator, replacing the data currently in x1. Adjust window as needed and you should get the following:



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

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

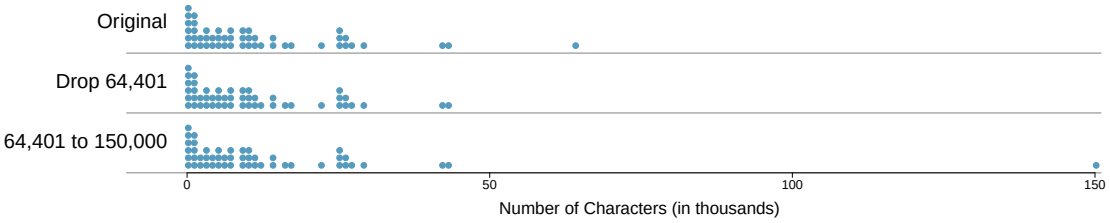


Figure 2.16: Dot plots of the original character count data and two modified data sets.

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

Figure 2.17: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change when extreme observations are present.

GUIDED PRACTICE 2.46

(G) (a) Which is more affected by extreme observations, the mean or median? Figure 2.17 may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?[35]

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 2.47

(E) The median and IQR do not change much under the three scenarios in Figure 2.17. Why might this be the case?

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

GUIDED PRACTICE 2.48

(G) 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?[36]

[35] (a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Guided Practice 2.46.

[36] 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.

2.2.7 Linear transformations of data

EXAMPLE 2.49

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

(E)

The original list has a mean of 3 and a standard deviation of 2. The new list: 10, 10, 50, 50 has a mean of 30 with a standard deviation of 20. Because all of the values were multiplied by 10, both the mean and the standard deviation were multiplied by 10. [37]

EXAMPLE 2.50

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

(E)

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

EXAMPLE 2.51

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

(E)

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

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

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

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

$$aX + b$$

Using the examples as a guide, we can solve this temperature-conversion problem. The mean was 27°C and the standard deviation was 3°C. To convert to Fahrenheit, we multiply all of the values by 9/5, which multiplies both the mean and the standard deviation by 9/5. Then we add 32 to all

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

of the values which adds 32 to the mean but does not change the standard deviation further.

$$\bar{x}_F = \frac{9}{5}\bar{x}_C + 32 \qquad \sigma_F = \frac{9}{5}\sigma_C$$
$$= \frac{5}{9}(27) + 32 \qquad = \frac{9}{5}(3)$$
$$= 80.6 \qquad = 5.4$$

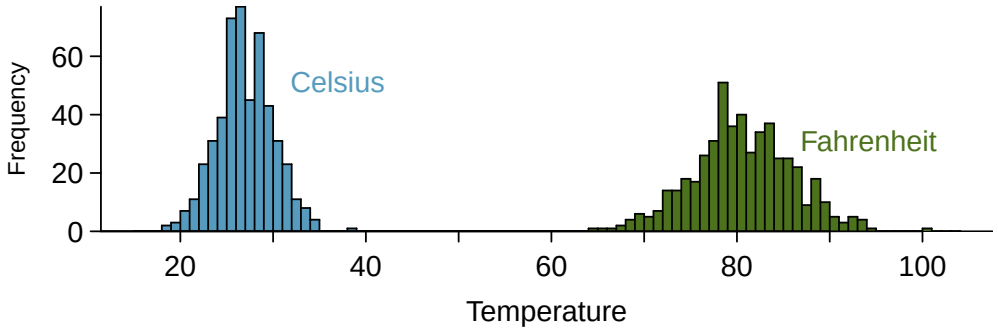


Figure 2.18: 500 temperatures shown in both Celsius and Fahrenheit.

ADDING SHIFTS THE VALUES, MULTIPLYING STRETCHES OR CONTRACTS THEM

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

EXAMPLE 2.52

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

(E)

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

For a more mathematical explanation of the IQR calculation, see the footnote.[38]

2.2.8 Comparing numerical data across groups

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

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

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

Median Income for 150 Counties, in $1000s

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

Figure 2.19: In this table, median household income (in $1000s) from a random sample of 100 counties that had population gains are shown on the left. Median incomes from a random sample of 50 counties that had no population gain are shown on the right.

There were 1,454 counties where the population increased from 2010 to 2017, and there were 1,672 counties with no gain (all but one were a loss). A random sample of 100 counties from the first group and 50 from the second group are shown in Figure 2.19 to give a better sense of some of the raw median income data.

The **side-by-side box plot** is a traditional tool for comparing across groups. An example is shown in the left panel of Figure 2.21, 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 2.21.

GUIDED PRACTICE 2.53

(G) Use the plots in Figure 2.21 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?[39]

COMPARING DISTRIBUTIONS

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

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

```
Population: Gain          Population: No Gain

                        | 2 |12
                        | 2 |79
                      4 | 3 |1234
              99999987555 | 3 |5555799
          444433322111000 | 4 |00111223333
99999888766666666555 5 | 4 |55666666789
44433322221111110000 | 5 |1112222
            887776666555 | 5 |6677
             33333222100 | 6 |01
                       9 | 6 |
                      42 | 7 |
                      85 | 7 |
                      21 | 8 |

Legend: 2 |1 = 21,000 median income
```

Figure 2.20: Back-to-back stem-and-leaf plot for median income, split by whether the count had a population gain or no gain.

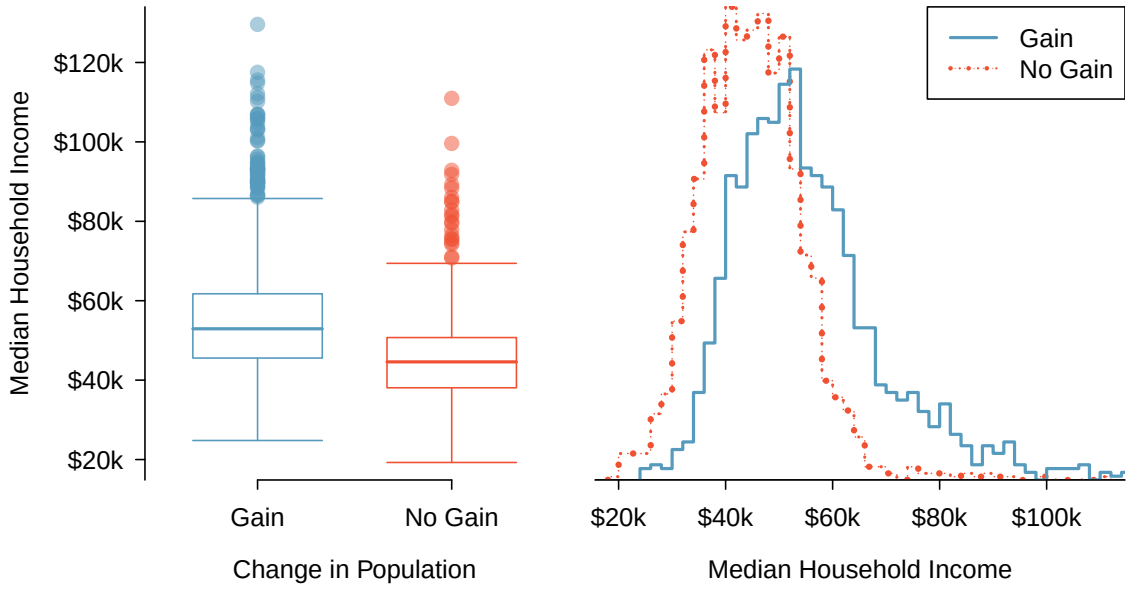


Figure 2.21: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_hh_income`, where the counties are split by whether or not there was a population gain from 2010 to 2017. Explore this data set on Tableau Public.

G **GUIDED PRACTICE 2.54**

What components of each plot in Figure 2.21 do you find most useful?[40]

G **GUIDED PRACTICE 2.55**

Do these graphs tell us about any association between income for the two groups?[41]

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

COMPARING DISTRIBUTIONS VERSUS LOOKING AT ASSOCIATION

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

2.2.9 Mapping data (special topic)

The `county` data set offers many numerical variables that we could plot using dot plots, scatterplots, or box plots, but these miss the true nature of the data. Rather, when we encounter geographic data, we should create an **intensity map**, where colors are used to show higher and lower values of a variable. Figures 2.22 and 2.23 shows intensity maps for poverty rate in percent (`poverty`), unemployment rate (`unemployment_rate`), homeownership rate in percent (`homeownership`), and median household income (`median_hh_income`). The color key indicates which colors correspond to which values. The intensity maps are not generally very helpful for getting precise values in any given county, but they are very helpful for seeing geographic trends and generating interesting research questions or hypotheses.

EXAMPLE 2.56

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

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

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

G **GUIDED PRACTICE 2.57**

What interesting features are evident in the `median_hh_income` intensity map in Figure 2.23(b)?[42]

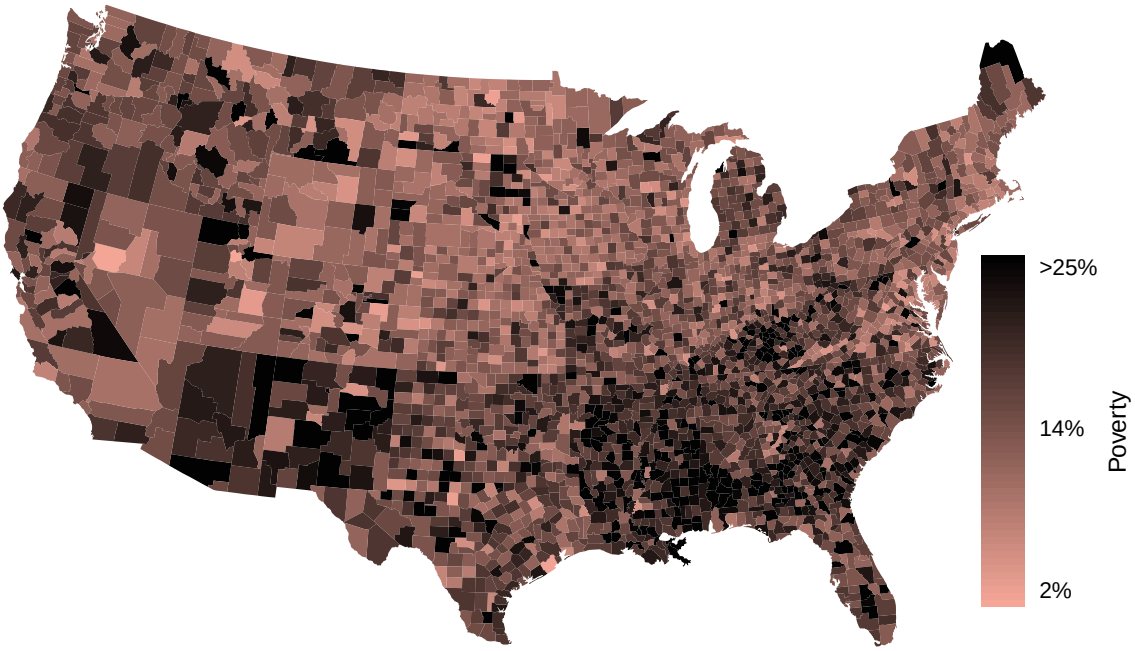
[40] Answers will vary. The parallel box plots are especially useful for comparing centers and spreads, while the hollow histograms are more useful for seeing distribution shape, skew, and groups of anomalies.

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

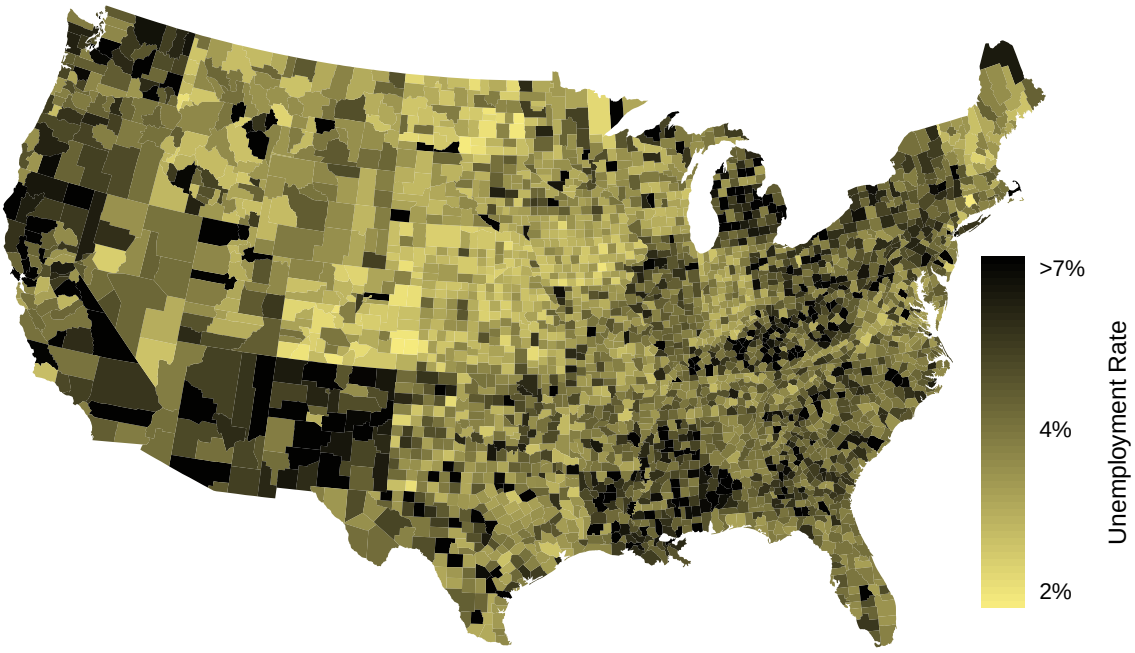
Section summary

- In this section we looked at univariate summaries, including two measures of **center** and three measures of **spread**.
- When **summarizing** or **comparing distributions**, always comment on center, spread, and shape. Also, mention outliers or gaps if applicable. Put descriptions in *context*, that is, identify the variable(s) being summarized by name and include relevant units. Remember: *Center, Spread, and Shape! In context!*
- **Mean** and **median** are measures of center. (A common mistake is to report **mode** as a measure of center. However, a mode can appear anywhere in a distribution.)
  - The **mean** is the sum of all the observations divided by the number of observations, n. $\bar{x} = \frac{1}{n}\sum x_i = \frac{\sum x_i}{n} = \frac{x_1+x_2+\ldots+x_n}{n}$
  - In an ordered data set, the **median** is the middle number when n is odd. When n is even, the median is the average of the two middle numbers.
- Because large values exert more "pull" on the mean, large values on the high end tend to increase the mean more than they increase the median. In a **right skewed** distribution, therefore, the mean is greater than the median. Analogously, in a **left skewed** distribution, the mean is less than the median. Remember: *The mean follows the tail! The skew is the tail!*
- **Standard deviation (SD)** and **Interquartile range (IQR)** are measures of spread. SD measures the typical spread from the mean, whereas IQR measures the spread of the middle 50% of the data.
  - To calculate the standard deviation, subtract the average from each value, square all those differences, add them up, divide by $n-1$, then take the square root. Note: The standard deviation is the square root of the variance.
    $s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$
  - The IQR is the difference between the third quartile Q_3 and the first quartile Q_1.
    $IQR = Q_3 - Q_1$
- **Range** is also sometimes used as a measure of spread. The range of a data set is defined as the difference between the maximum value and the minimum value, i.e. $max - min$.
- **Outliers** are observations that are extreme relative to the rest of the data. Two rules of thumb for identifying observations as outliers are:
  - more than 2 standard deviations above or below the mean
  - more than $1.5 \times IQR$ below Q_1 or above Q_3
- Mean and SD are sensitive to outliers. Median and IQR are more robust and less sensitive to outliers.
- A **Z-score** represents the number of standard deviations a value in a data set is above or below the mean. To calculate a Z-score use: $Z = \frac{x - \text{mean}}{SD}$.
- *Z-scores do not depend on units.* When looking at distributions with different units or different standard deviations, Z-scores are useful for comparing how far values are away from the mean (relative to the distribution of the data).
- **Linear transformations of data**. Adding a constant to every value in a data set shifts the mean but does not affect the standard deviation. Multiplying the values in a data set by a constant will multiply the mean and the standard deviation by that constant, except that the standard deviation must always remain positive.

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

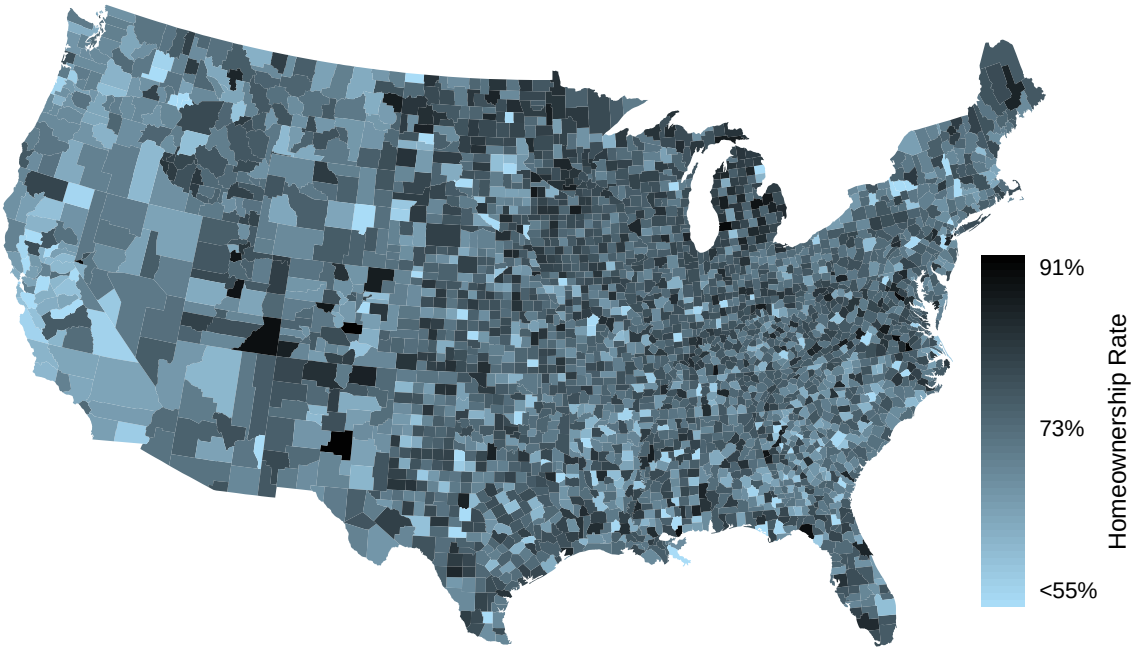


(a)

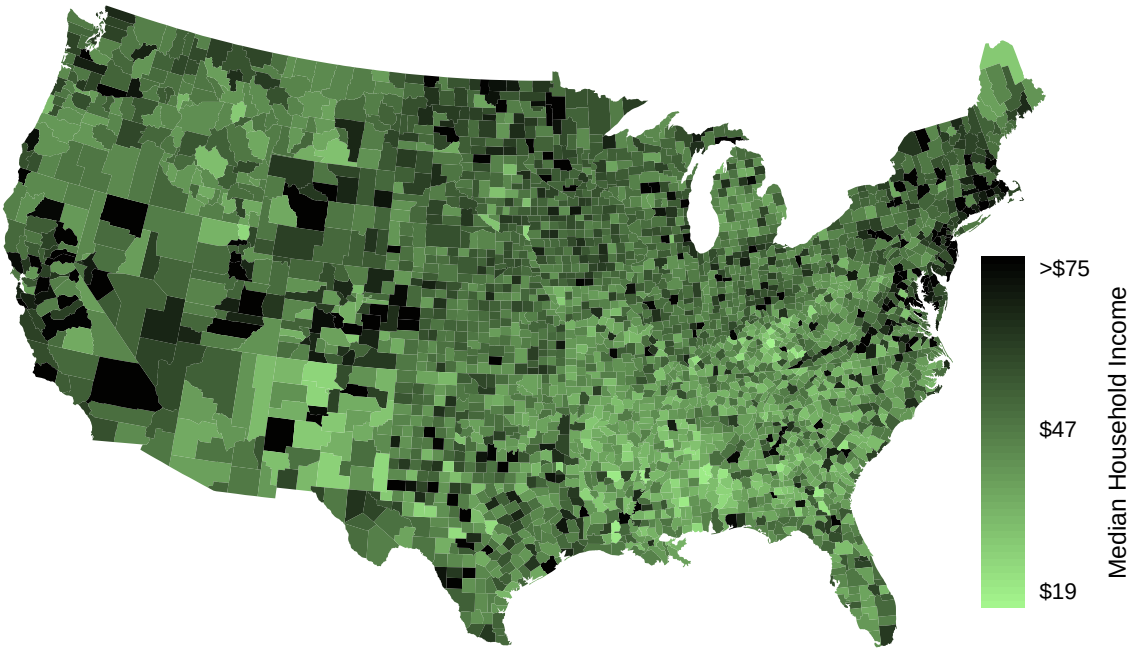


(b)

Figure 2.22: (a) Intensity map of poverty rate (percent). (b) Intensity map of the unemployment rate (percent). Explore dozens of intensity maps using American Community Survey data on Tableau Public.



(a)



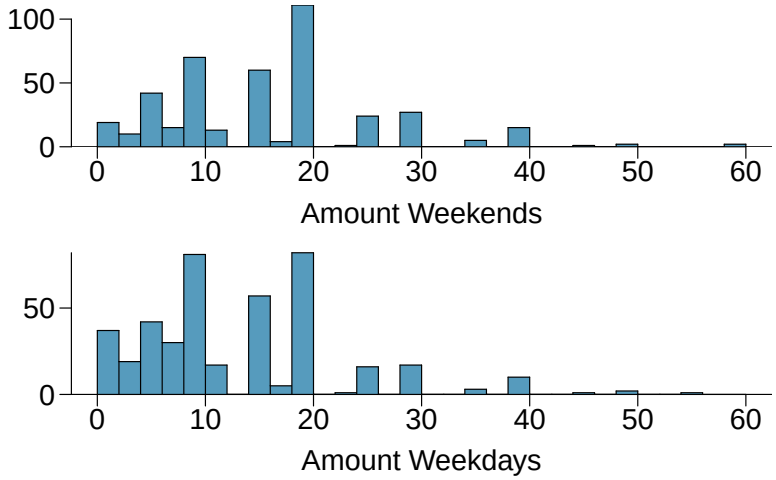
(b)

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

- **Box plots** do not show the *distribution* of a data set in the way that histograms do. Rather, they provide a visual depiction of the **5-number summary**, which consists of: min, Q_1, Q_2, Q_3, max. While a box plot does not indicate modes, it can show skew and outliers.

Exercises

2.7 Smoking habits of UK residents, Part I. A survey was conducted to study the smoking habits of UK residents. The histograms below display the distributions of the number of cigarettes smoked on weekdays and weekends, and they exclude data from people who identified themselves as non-smokers. Describe the two distributions and compare them.[43]



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

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

Draw a histogram of these data and describe the distribution.

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

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

(a) Find the mean amount of cigarettes smoked on weekdays and weekends by these 5 respondents.

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

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

(a) Calculate the mean for these data.

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

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

[43] **data:smoking**.

2.12 Medians and IQRs. For each part, compare distributions (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning.

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

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

(c) (1) 1, 2, 3, 4, 5
(2) 6, 7, 8, 9, 10

(d) (1) 0, 10, 50, 60, 100
(2) 0, 100, 500, 600, 1000

2.13 Means and SDs. For each part, compare distributions (1) and (2) based on their means and standard deviations. You do not need to calculate these statistics; simply state how the means and the standard deviations compare. Make sure to explain your reasoning. *Hint:* It may be useful to sketch dot plots of the distributions.

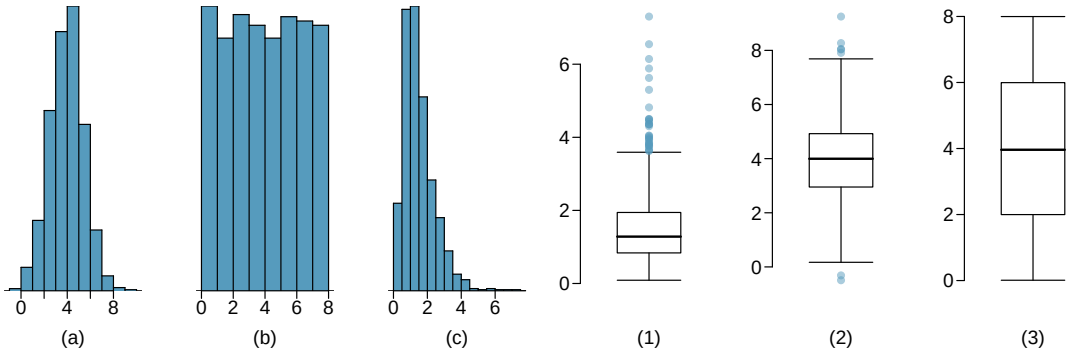
(a) (1) 3, 5, 5, 5, 8, 11, 11, 11, 13
(2) 3, 5, 5, 5, 8, 11, 11, 11, 20

(b) (1) -20, 0, 0, 0, 15, 25, 30, 30
(2) -40, 0, 0, 0, 15, 25, 30, 30

(c) (1) 0, 2, 4, 6, 8, 10
(2) 20, 22, 24, 26, 28, 30

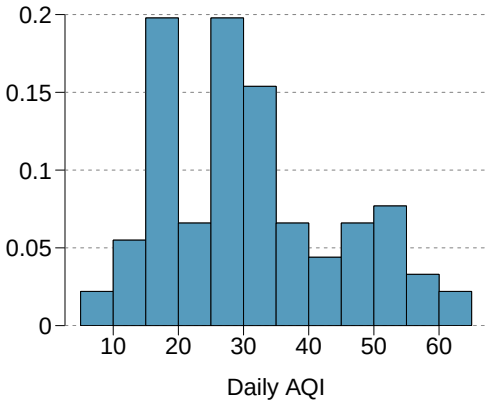
(d) (1) 100, 200, 300, 400, 500
(2) 0, 50, 300, 550, 600

2.14 Mix-and-match. Describe the distribution in the histograms below and match them to the box plots.



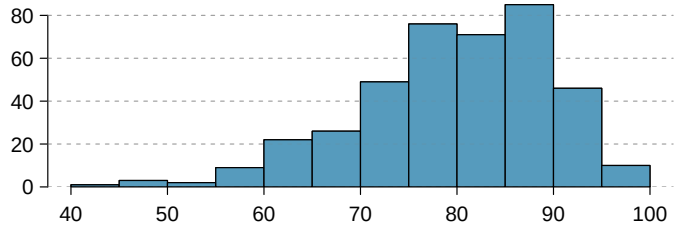
2.15 Air quality. Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of 91 days in 2011 in Durham, NC. The relative frequency histogram below shows the distribution of the AQI values on these days.[44]

(a) Estimate the median AQI value of this sample.
(b) Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.
(c) Estimate Q1, Q3, and IQR for the distribution.
(d) Would any of the days in this sample be considered to have an unusually low or high AQI? Explain your reasoning.

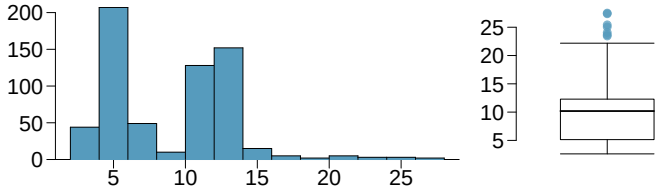


[44] **data:durhamAQI:2011**.

2.16 Median vs. mean. Estimate the median for the 400 observations shown in the histogram, and note whether you expect the mean to be higher or lower than the median.



2.17 Histograms vs. box plots. Compare the two plots below. What characteristics of the distribution are apparent in the histogram and not in the box plot? What characteristics are apparent in the box plot but not in the histogram?



2.18 Facebook friends. Facebook data indicate that 50% of Facebook users have 100 or more friends, and that the average friend count of users is 190. What do these findings suggest about the shape of the distribution of number of friends of Facebook users?[45]

2.19 Distributions and appropriate statistics, Part I. For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Number of pets per household.

(b) Distance to work, i.e. number of miles between work and home.

(c) Heights of adult males.

2.20 Distributions and appropriate statistics, Part II. For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

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

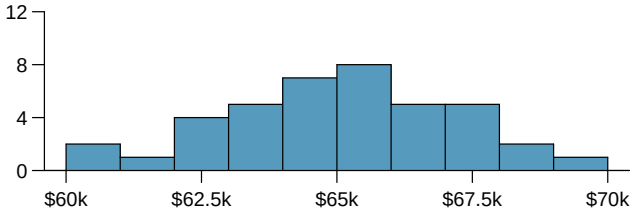
(b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

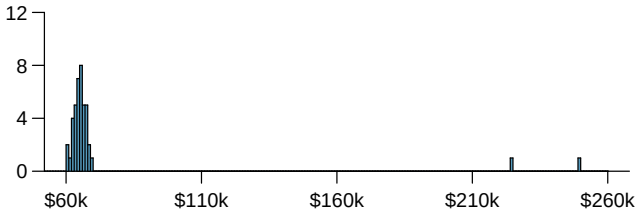
(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than all the other employees.

[45] **Backstrom:2011**.

2.21 Income at the coffee shop. The first histogram below shows the distribution of the yearly incomes of 40 patrons at a college coffee shop. Suppose two new people walk into the coffee shop: one making $225,000 and the other $250,000. The second histogram shows the new income distribution. Summary statistics are also provided.



(1)



(2)

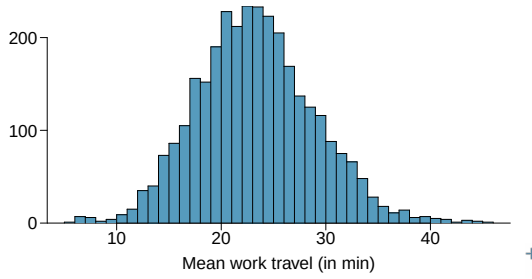
	(1)	(2)
n	40	42
Min.	60,680	60,680
1st Qu.	63,620	63,710
Median	65,240	65,350
Mean	65,090	73,300
3rd Qu.	66,160	66,540
Max.	69,890	250,000
SD	2,122	37,321

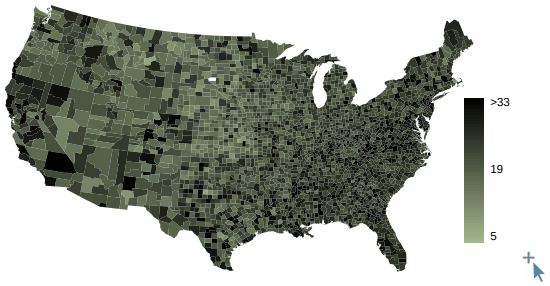
(a) Would the mean or the median best represent what we might think of as a typical income for the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

(b) Would the standard deviation or the IQR best represent the amount of variability in the incomes of the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

2.22 Midrange. The *midrange* of a distribution is defined as the average of the maximum and the minimum of that distribution. Is this statistic robust to outliers and extreme skew? Explain your reasoning

2.23 Commute times. The US census collects data on time it takes Americans to commute to work, among many other variables. The histogram below shows the distribution of average commute times in 3,142 US counties in 2017.

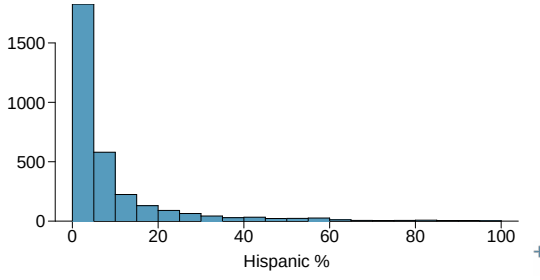


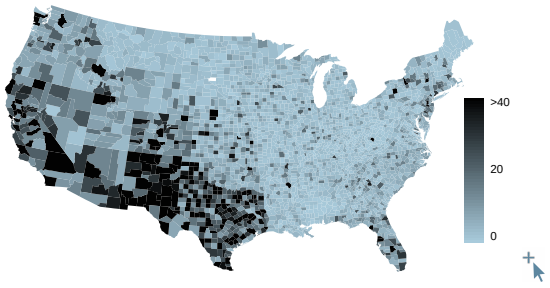


(a) Describe the distribution of average commute times for counties in the US.

(b) Describe the spatial distribution of commuting times using the map provided.

2.24 Hispanic/Latinx population. The US census collects data on race and ethnicity of Americans, among many other variables. The histogram below shows the distribution of the percentage of the population that is Hispanic/Latinx in 3,142 counties in the US in 2017.





(a) Describe the distribution of percent of population that is Hispanic/Latinx for counties in the US.

(b) What features of the distribution of the Hispanic/Latinx population in US counties are apparent in the map but not in the histogram? What features are apparent in the histogram but not the map?

2.3 Normal distribution

What proportion of adults have systolic blood pressure above 140? What is the probability of getting more than 250 heads in 400 tosses of a fair coin? If the average weight of a piece of carry-on luggage is 11 pounds, what is the probability that 200 random carry on pieces will weigh more than 2500 pounds? If 55% of a population supports a certain candidate, what is the probability that she will have less than 50% support in a random sample of size 200?

There is one distribution that can help us answer all of these questions. Can you guess what it is? That's right – it's the normal distribution.

Learning objectives

1. Use Z-scores and the standard normal model to approximate a distribution where appropriate.
2. Find probabilities and percentiles using the normal approximation.
3. Find the value that corresponds to a given percentile when the distribution is approximately normal.

2.3.1 Normal distribution model

Among all the distributions we see in practice, one is overwhelmingly the most common. The symmetric, unimodal, bell curve is ubiquitous throughout statistics. Indeed it is so common, that people often know it as the **normal curve** or **normal distribution**.[46] A normal curve is shown in Figure 2.24.

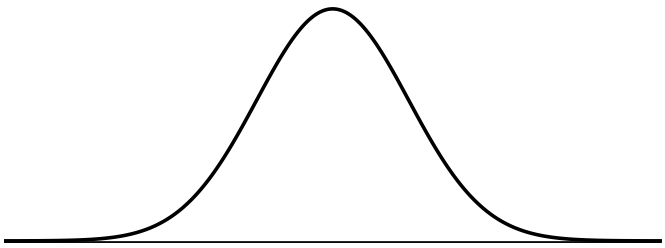


Figure 2.24: A normal curve.

The **normal distribution** always describes a symmetric, unimodal, bell-shaped curve. However, these curves can look different depending on the details of the model. Specifically, the normal distribution model can be adjusted using two parameters: mean and standard deviation. As you can probably guess, changing the mean shifts the bell curve to the left or right, while changing the standard deviation stretches or constricts the curve.

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

Figure 2.25 shows the normal distribution with mean 0 and standard deviation 1 in the left panel and the normal distributions with mean 19 and standard deviation 4 in the right panel. Figure 2.26 shows these distributions on the same axis.

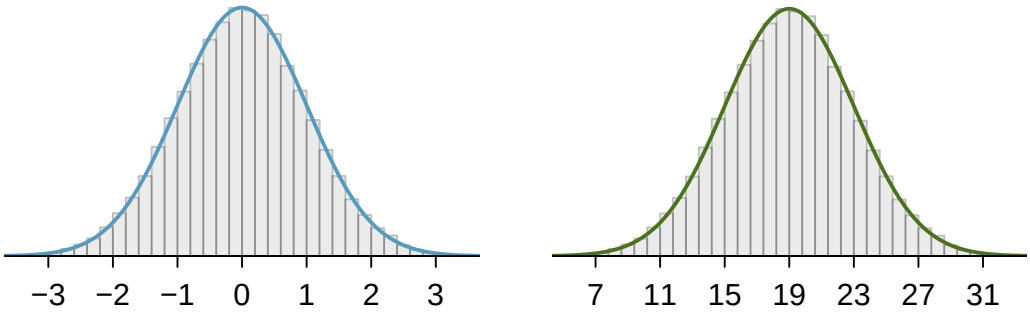


Figure 2.25: Both curves represent the normal distribution. However, they differ in their center and spread.

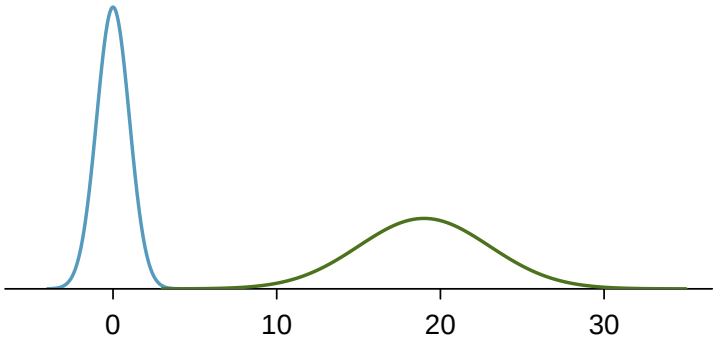


Figure 2.26: The normal distributions shown in Figure 2.25 but plotted together and on the same scale.

Because the mean and standard deviation describe a normal distribution exactly, they are called the distribution's **parameters**. The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard normal distribution**..

NORMAL DISTRIBUTION FACTS

Many variables are nearly normal, but none are exactly normal. The normal distribution, while never perfect, provides very close approximations for a variety of scenarios. We will use it to model data as well as probability distributions.

2.3.2 Using the normal distribution to approximate empirical distributions

We often want to put data onto a standardized scale, which can make comparisons more reasonable.

EXAMPLE 2.58

Figure 2.27 shows the mean and standard deviation for total scores on the SAT and ACT. The distribution of SAT and ACT scores are both nearly normal. Suppose Ann scored 1300 on her SAT and Tom scored 24 on his ACT. Who performed better?

As we saw in section 2.2.3, we can use Z-scores to compare observations from different distributions. Using Ann's SAT score, 1300 , along with the SAT mean and SD, we can find Ann's Z-score.

(E)

$$Z_{\text{Ann}} = \frac{x_{\text{Ann}} - \mu_{\text{SAT}}}{\sigma_{\text{SAT}}} = \frac{1300 - 1100}{200} = 1$$

Similarly, using Tom's ACT score, 24, along with the ACT mean and SD we can find his Z-score.

$$Z_{Tom} = \frac{x_{\text{Tom}} - \mu_{\text{ACT}}}{\sigma_{\text{ACT}}} = \frac{24 - 21}{6} = 0.5$$

Because Ann's score was 1 standard deviation above the mean, while Tom's score was 0.5 standard deviations above the mean, we can say that Ann did better than Tom.

	SAT	ACT
Mean	1100	21
SD	200	6

Figure 2.27: Mean and standard deviation for the SAT and ACT.

Assuming that both the SAT and ACT distributions are nearly normally distributed, what percent of test takers scored lower than Ann? What percent scored lower than Tom? To answer these question exactly, we would need all of the data. However, if we use the information that SAT and ACT distributions are nearly normal, we can estimate these percents. Figure 2.28 shows these distributions modeled with a normal curve. If we can find the percent of the normal curve that is to the left of Ann's score, we could use that percent as our estimate of the percent of the data points that are smaller than Ann's score. We call this process *normal approximation*. The steps are:

1. First verify that the distribution can be reasonably modeled with a normal distribution.
2. Convert value or values of interest to Z-scores.
3. Find the relevant area/percent under the standard normal curve.

We use the area/percent that we find from the normal curve as our *estimate* of the desired percent.

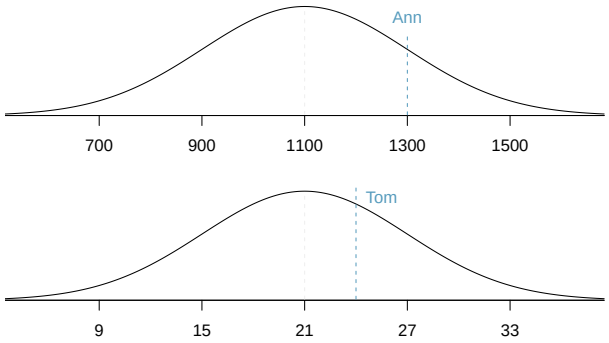


Figure 2.28: Ann's and Tom's scores shown with the distributions of SAT and ACT scores.

2.3.3 Finding areas under the normal curve

It's very useful in statistics to be able to identify areas of distributions, especially tail areas. For instance, what percent of people have an SAT score below Ann's score of 1300? This is the same as Ann's **percentile**. We previously determined that a score of 1300 corresponds to a Z-score of 1 and that SAT scores are approximately normally distributed. We can visualize such a tail area by sketching a normal curve and shading everything below $Z = 1$ as shown in Figure 2.29.

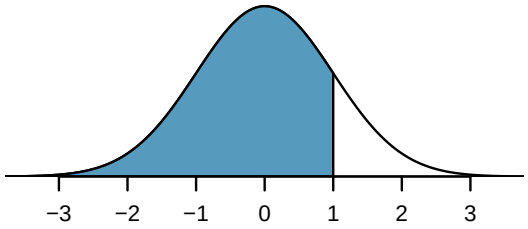


Figure 2.29: The area to the left of the Z-score represents the percentile of the observation.

There are many techniques for finding this area, and we'll discuss three of the options.

1. The most common approach in practice is to use statistical software. For example, in the program **R**, we could find the area shown in Figure 2.29 using the following command, which takes in the Z-score of 1 and returns the lower tail area:

   ```
   > pnorm(1)
   [1] 0.8413447
   ```

   Using the online Desmos calculator, we could do: `normaldist( )`, check the "Find Cumulative Probability (CDF)" box and set Max to 1.

   According to these calculation, the area shaded that is below $Z = 1$ is 0.841, so we estimate that 84.1% of SAT test takers score below 1300 and that Ann is at the 84^{th} percentile.

   There are many other software options, such as Python or SAS; even spreadsheet programs such as Excel and Google Sheets support these calculations.

2. A common strategy in classrooms is to use a graphing calculator, such as a TI or Casio calculator. Instructions for finding areas of a normal distribution using these calculators are provided in Section 2.3.6.

3. The last option for finding tail areas is to use what's called a **probability table**; these are occasionally used in classrooms but rarely in practice. Appendix C.2 contains such a table and a guide for how to use it.

We will solve normal distribution problems in this section by always first finding the Z-score. The reason is that we will encounter close parallels called test statistics beginning in Chapter **??**; these are, in many instances, an equivalent of a Z-score.

Readers may find it helpful to familiarize themselves with one of the options above before continuing on to the applications that follow.

2.3.4 Normal probability examples

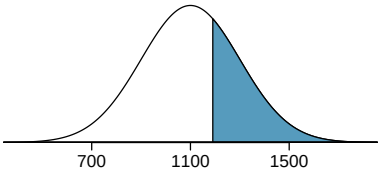
CumulativeCombined SAT scores are approximated well by a normal model with mean 1100 and standard deviation 200.

EXAMPLE 2.59

What is the probability that a randomly selected SAT taker scores at least 1190 on the SAT?

The probability that a randomly selected SAT taker scores at least 1190 on the SAT is equivalent to the proportion of all SAT takers that score at least 1190 on the SAT. First, always draw and label a picture of the normal distribution. (Drawings need not be exact to be useful.) We are interested in the probability that a randomly selected score will be above 1190, so we shade this upper tail:

E



The picture shows the mean and the values at 2 standard deviations above and below the mean. The simplest way to find the shaded area under the curve makes use of the Z-score of the cutoff value. With $\mu = 1100$, $\sigma = 200$, and the cutoff value $x = 1190$, the Z-score is computed as

$$Z = \frac{x - \mu}{\sigma} = \frac{1190 - 1100}{200} = \frac{90}{200} = 0.45$$

Next, we want to find the area under the normal curve to right of $Z = 0.45$. Using technology, we find $P(Z > 0.45) = 0.3264$. The probability that a randomly selected score is at least 1190 on the SAT is 0.3264.

ALWAYS DRAW A PICTURE FIRST, AND FIND THE Z-SCORE SECOND

For any normal probability situation, *always always always* draw and label the normal curve and shade the area of interest first. The picture will provide an estimate of the probability.

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

GUIDED PRACTICE 2.60

G

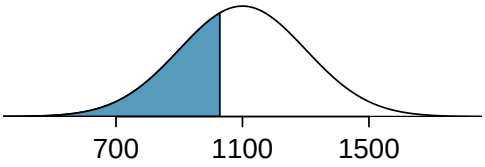
If the probability that a randomly selected score is at least 1190 is 0.3264, what is the probability that the score is less than 1190? Draw the normal curve representing this exercise, shading the lower region instead of the upper one.[47]

[47] We found the probability in Example 2.59: 0.6736. A picture for this exercise is represented by the shaded area below "0.6736" in Example 2.59.

EXAMPLE 2.61

Edward earned a 1030 on his SAT. What is his percentile?

First, a picture is needed. Edward's percentile is the proportion of people who do not get as high as a 1030. These are the scores to the left of 1030.



Identifying the mean $\mu = 1100$, the standard deviation $\sigma = 200$, and the cutoff for the tail area $x = 1030$ makes it easy to compute the Z-score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1030 - 1100}{200} = -0.35$$

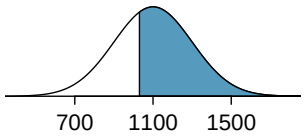
Using technology we find that $P(Z < -0.35) = 0.3632$. Edward is at the 36^{th} percentile.

GUIDED PRACTICE 2.62

Use the results of Example 2.61 to compute the proportion of SAT takers who did better than Edward. Also draw a new picture.[48]

The last several problems have focused on finding the probability or percentile for a particular observation. It is also possible to identify the value corresponding to a particular percentile.

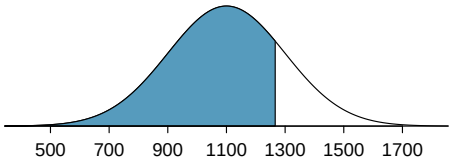
[48] If Edward did better than 36% of SAT takers, then about 64% must have done better than him.



EXAMPLE 2.63

Carlos believes he can get into his preferred college if he scores at least in the 80th percentile on the SAT. What score should he aim for?

Here, we are given a percentile rather than a Z-score, so we work backwards. As always, first draw the picture.



E

We want to find the observation that corresponds to the 80th percentile. First, we find the Z-score associated with the 80th percentile. Using technology, we find that $P(Z < 0.84) = 0.80$. In any normal distribution, a value with a Z-score of 0.84 will be at the 80th percentile. Once we have the Z-score, we work backwards to find x.

$$\begin{aligned} Z &= \frac{x - \mu}{\sigma} \\ 0.84 &= \frac{x - 1100}{200} \\ 0.84 \times 200 + 1100 &= x \\ x &= 1268 \end{aligned}$$

The 80th percentile on the SAT corresponds to a score of 1268.

G

GUIDED PRACTICE 2.64

Imani scored at the 72nd percentile on the SAT. What was her SAT score?[49]

[reworded]

IF THE DATA ARE NOT NEARLY NORMAL, DON'T USE NORMAL APPROXIMATION

Before using the normal approximation method, verify that the data or distribution is approximately normal. If it is not, the normal approximation will give incorrect results. Also remember: all answers based on normal approximations are approximations and are not exact.

Finally, we should observe that it is possible for a normal random variable to fall 4, 5, or even more standard deviations from the mean. The probability of being further than 4 standard deviations from the mean is about 1-in-15,000. For 5 and 6 standard deviations, it is about 1-in-2 million and 1-in-500 million, respectively. However, while the tails of the normal distribution extend infinitely in either direction, our data sets are finite and normal approximation in the extreme tails is unlikely to be very accurate, even for bell-shaped data sets.

[49] First, draw a picture! The closest percentile in the table to 0.72 is 0.7190, which corresponds to $Z = 0.58$. Next, set up the Z-score formula and solve for x: $0.58 = \frac{x-1100}{200} \rightarrow x = 1216$. Imani scored 1216.

2.3.5 Evaluating the normal approximation (special topic)

It is important to remember normality is always an approximation. Testing the appropriateness of the normal assumption is a key step in many data analyses.

The distribution of heights of US males is well approximated by the normal model. We are interested in proceeding under the assumption that the data are normally distributed, but first we must check to see if this is reasonable.

There are two visual methods for checking the assumption of normality that can be implemented and interpreted quickly. The first is a simple histogram with the best fitting normal curve overlaid on the plot, as shown in the left panel of Figure 2.30. 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**,[50] shown in the right panel of Figure 2.30. The closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model.

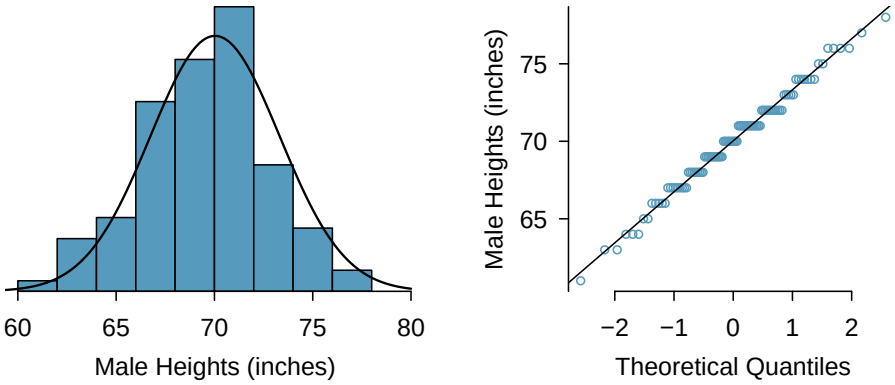


Figure 2.30: A sample of 100 male heights. The observations are rounded to the nearest whole inch, explaining why the points appear to jump in increments in the normal probability plot.

[check placement of figures]

EXAMPLE 2.65

Consider all NBA players from the 2018-2019 season presented in Figure 2.31.[51] Based on the graphs, are NBA player heights normally distributed?

(E)

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". NBA player heights do not appear to come from a normal distribution.

GUIDED PRACTICE 2.66

(G) Figure 2.32 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?[52]

[50] Also commonly called a quantile-quantile plot.

[51] These data were collected from www.nba.com.

[52] 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 right skewed. The second plot shows the opposite features, and this distribution is left skewed.

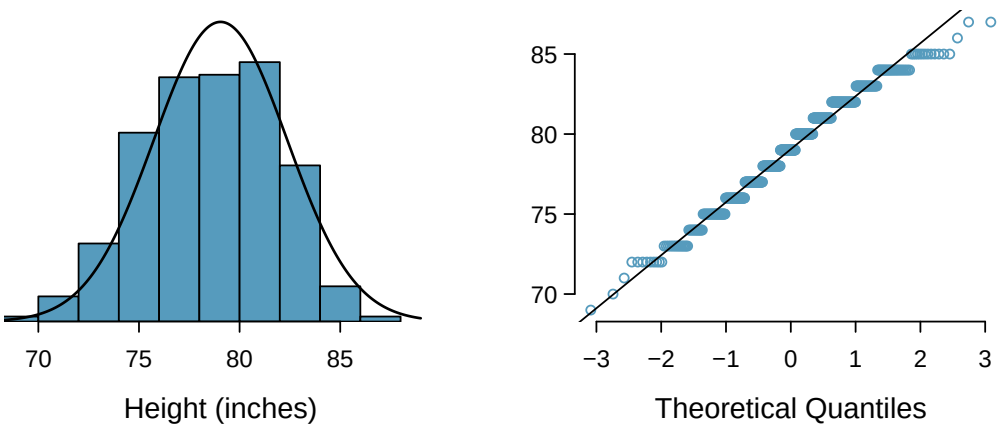


Figure 2.31: Histogram and normal probability plot for the NBA heights from the 2018-2019 season.

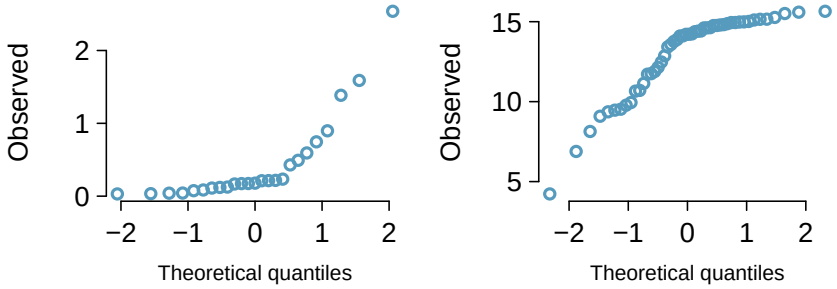
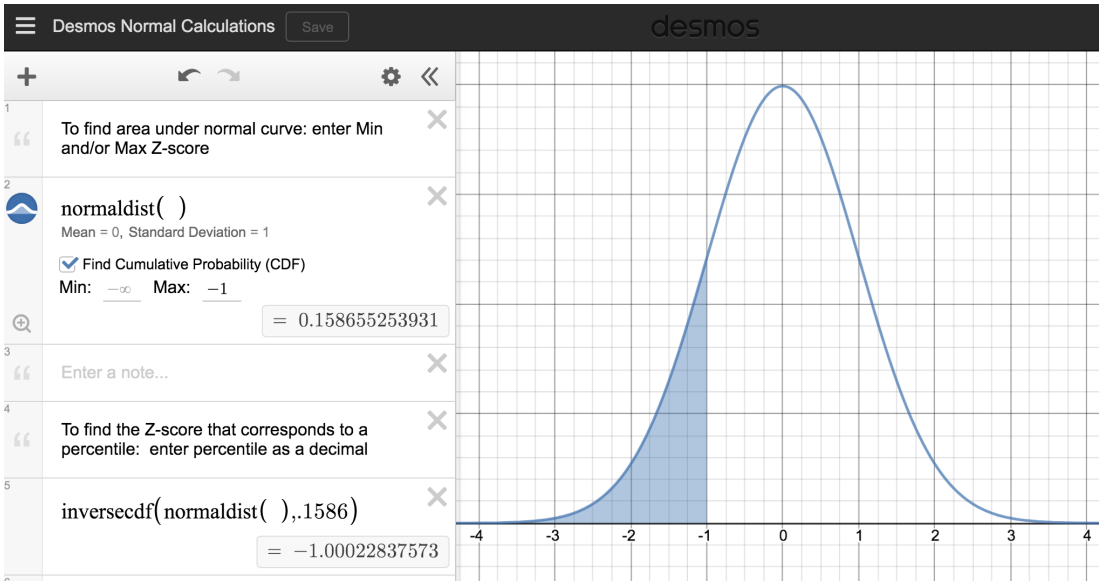


Figure 2.32: Normal probability plots for Guided Practice 2.66.

2.3.6 Technology: finding normal probabilities

Get started quickly with this Desmos calculator.



Calculator instructions

TI-84: FINDING AREA UNDER THE NORMAL CURVE

Use 2ND VARS, normalcdf to find an area/proportion/probability between two Z-scores or to the left or right of a Z-score.

1. Choose 2ND VARS (i.e. DISTR).
2. Choose 2:normalcdf.
3. Enter the lower (left) Z-score and the upper (right) Z-score.
   - If finding just a lower tail area, set lower to -5.
   - If finding just an upper tail area, set upper to 5.
4. Leave μ as 0 and σ as 1.
5. Down arrow, choose Paste, and hit ENTER.

TI-83: Do steps 1-2, then enter the lower bound and upper bound separated by a comma, e.g. normalcdf(2, 5), and hit ENTER.

CASIO FX-9750GII: FINDING AREA UNDER THE NORMAL CURVE

1. Navigate to STAT (MENU, then hit 2).
2. Select DIST (F5), then NORM (F1), and then Ncd (F2).
3. If needed, set Data to Variable (Var option, which is F2).
4. Enter the Lower Z-score and the Upper Z-score. Set σ to 1 and μ to 0.
   - If finding just a lower tail area, set Lower to -5.
   - For an upper tail area, set Upper to 5.
5. Hit EXE, which will return the area probability (p) along with the Z-scores for the lower and upper bounds.

GUIDED PRACTICE 2.67

Use a calculator or software 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.[53]

GUIDED PRACTICE 2.68

Find the area under the normal curve between -1.5 and 1.5. [54]

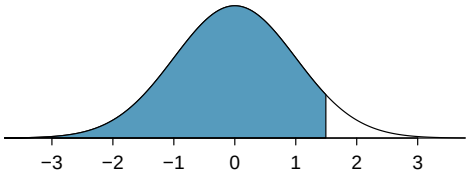
[53] To find the area between $Z = -1$ and $Z = 1$, let lower bound be -1 and upper bound be 1. We find that $P(-1 < Z < 1) = 0.6827$. Similarly, $P(-2 < Z < 2) = 0.9545$ and $P(-3 < Z < 3) = 0.9973$.

[54] Lower bound is -1.5 and upper bound is 1.5. The area under the normal curve between -1.5 and 1.5 = $P(-1.5 < Z < 1.5) = 0.866$. Note that is not simply the average of 0.6827 and 0.9545, as the normal curve is not a rectangle.

EXAMPLE 2.69

Use a calculator to determine what percentile corresponds to a Z-score of 1.5.[55]

E



To find an area under the normal curve using a calculator, first identify a lower bound and an upper bound. We want all of the area to the left of 1.5, so the lower bound should be -∞. However, the area under the curve is negligible when Z is smaller than -5, so we will use -5 as the lower bound. Using a lower bound of -5 and an upper bound of 1.5, we get $P(Z < 1.5) = 0.933$.

G

GUIDED PRACTICE 2.70

Find the area under the normal curve to right of $Z = 2$. [56]

TI-84: FIND A Z-SCORE THAT CORRESPONDS TO A PERCENTILE

Use `2ND VARS`, `invNorm` to find the Z-score that corresponds to a given percentile.

1. Choose `2ND VARS` (i.e. `DISTR`).
2. Choose `3:invNorm`.
3. Let `Area` be the percentile as a decimal (the area to the left of desired Z-score).
4. Leave μ as `0` and σ as `1`.
5. Down arrow, choose `Paste`, and hit `ENTER`.

TI-83: Do steps 1-2, then enter the percentile as a decimal, e.g. `invNorm(.40)`, then hit `ENTER`.

CASIO FX-9750GII: FIND A Z-SCORE THAT CORRESPONDS TO A PERCENTILE

1. Navigate to `STAT` (`MENU`, then hit `2`).
2. Select `DIST` (`F5`), then `NORM` (`F1`), and then `InvN` (`F3`).
3. If needed, set `Data` to `Variable` (`Var` option, which is `F2`).
4. Decide which tail area to use (`Tail`), the tail area (`Area`), and then enter the σ and μ values.
5. Hit `EXE`.

EXAMPLE 2.71

Use a calculator to find the Z-score that corresponds to the 40th percentile.

E

Letting area be 0.40, a calculator gives -0.253. This means that $Z = -0.253$ corresponds to the 40th percentile, that is, $P(Z < -0.253) = 0.40$.

G

GUIDED PRACTICE 2.72

Find the Z-score such that 20 percent of the area is to the right of that Z-score.[57]

[55] normalcdf gives the result without drawing the graph. To draw the graph, do 2nd VARS, DRAW, 1:ShadeNorm. However, beware of errors caused by other plots that might interfere with this plot.

[56] Now we want to shade to the right. Therefore our lower bound will be 2 and the upper bound will be +5 (or a number bigger than 5) to get $P(Z > 2) = 0.023$.

Section summary

- A **Z-score** represents the number of standard deviations a value in a data set is above or below the mean. To calculate a Z-score use: $Z = \frac{x - \text{mean}}{SD}$.
- The **normal distribution** is the most commonly used distribution in Statistics. Many distribution are approximately normal, but none are exactly normal.
- The empirical rule (68-95-99.7 Rule) comes from the normal distribution. The closer a distribution is to normal, the better this rule will hold.
- It is often useful to use the standard normal distribution, which has mean 0 and SD 1, to approximate a discrete histogram. There are two common types of **normal approximation problems**, and for each a key step is to find a Z-score.

  A: *Find the percent or probability of a value greater/less than a given x-value.*
  1. Verify that the distribution of interest is approximately normal.
  2. Calculate the Z-score. Use the provided population mean and SD to standardize the given x-value.
  3. Use a calculator function (e.g. `normcdf` on a TI) or other technology to find the area under the normal curve to the right/left of this Z-score; this is the *estimate* for the percent/probability.

  B: *Find the x-value that corresponds to a given percentile.*
  1. Verify that the distribution of interest is approximately normal.
  2. Find the Z-score that corresponds to the given percentile (using, for example, `invNorm` on a TI).
  3. Use the Z-score along with the given mean and SD to solve for the x-value.

[57] If 20% of the area is the right, then 80% of the area is to the left. Letting area be 0.80, we get $Z = 0.841$.

Exercises

2.25 Area under the curve, Part I. What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z < -1.35$ (b) $Z > 1.48$ (c) $-0.4 < Z < 1.5$ (d) $|Z| > 2$

2.26 Area under the curve, Part II. What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z > -1.13$ (b) $Z < 0.18$ (c) $Z > 8$ (d) $|Z| < 0.5$

2.27 GRE scores, Part I. Sophia who took the Graduate Record Examination (GRE) scored 160 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distributions are nearly normal.

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

(b) What do these Z-scores tell you?

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

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

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

(f) Explain why simply comparing raw scores from the two sections could lead to an incorrect conclusion as to which section a student did better on.

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

2.28 Triathlon times, Part I. In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the *Men, Ages 30 - 34* group while Mary competed in the *Women, Ages 25 - 29* group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

- The finishing times of the *Men, Ages 30 - 34* group has a mean of 4313 seconds with a standard deviation of 583 seconds.
- The finishing times of the *Women, Ages 25 - 29* group has a mean of 5261 seconds with a standard deviation of 807 seconds.
- The distributions of finishing times for both groups are approximately Normal.

Remember: a better performance corresponds to a faster finish.

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

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

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

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

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

2.29 GRE scores, Part II. In Exercise 2.27 we saw two distributions for GRE scores: $N(\mu = 151, \sigma = 7)$ for the verbal part of the exam and $N(\mu = 153, \sigma = 7.67)$ for the quantitative part. Use this information to compute each of the following:

(a) The score of a student who scored in the 80^{th} percentile on the Quantitative Reasoning section.

(b) The score of a student who scored worse than 70% of the test takers in the Verbal Reasoning section.

2.30 Triathlon times, Part II. In Exercise 2.28 we saw two distributions for triathlon times: $N(\mu = 4313, \sigma = 583)$ for *Men, Ages 30 - 34* and $N(\mu = 5261, \sigma = 807)$ for the *Women, Ages 25 - 29* group. Times are listed in seconds. Use this information to compute each of the following:

(a) The cutoff time for the fastest 5% of athletes in the men's group, i.e. those who took the shortest 5% of time to finish.

(b) The cutoff time for the slowest 10% of athletes in the women's group.

2.31 LA weather, Part I.  The average daily high temperature in June in LA is 77°F with a standard deviation of 5°F. Suppose that the temperatures in June closely follow a normal distribution.

(a) What is the probability of observing an 83°F temperature or higher in LA during a randomly chosen day in June?

(b) How cool are the coldest 10% of the days (days with lowest average high temperature) during June in LA?

2.32 CAPM. The Capital Asset Pricing Model (CAPM) is a financial model that assumes returns on a portfolio are normally distributed. Suppose a portfolio has an average annual return of 14.7% (i.e. an average gain of 14.7%) with a standard deviation of 33%. A return of 0% means the value of the portfolio doesn't change, a negative return means that the portfolio loses money, and a positive return means that the portfolio gains money.

(a) What percent of years does this portfolio lose money, i.e. have a return less than 0%?

(b) What is the cutoff for the highest 15% of annual returns with this portfolio?

2.33 LA weather, Part II. Exercise 2.31 states that average daily high temperature in June in LA is 77°F with a standard deviation of 5°F, and it can be assumed that they to follow a normal distribution. We use the following equation to convert °F (Fahrenheit) to °C (Celsius):

$$C = (F - 32) \times \frac{5}{9}.$$

(a) What is the probability of observing a 28°C (which roughly corresponds to 83°F) temperature or higher in June in LA? Calculate using the °C model from part (a).

(b) Did you get the same answer or different answers in part (b) of this question and part (a) of Exercise 2.31? Are you surprised? Explain.

(c) Estimate the IQR of the temperatures (in °C) in June in LA.

2.34 Find the SD. Cholesterol levels for women aged 20 to 34 follow an approximately normal distribution with mean 185 milligrams per deciliter (mg/dl). Women with cholesterol levels above 220 mg/dl are considered to have high cholesterol and about 18.5% of women fall into this category. Find the standard deviation of this distribution.

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

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

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

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

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

The mean height is 61.52 inches with a standard deviation of 4.58 inches. Use this information to determine if the heights approximately follow the 68-95-99.7% Rule.

2.4 Considering categorical data

How do we visualize and summarize categorical data? In this section, we will introduce tables and other basic tools for categorical data that are used throughout this book and will answer the following questions:

- Based on the `loan50` data, is there an assocation between the categorical variables of home-ownership and application type (individual, joint)?
- Using the `email50` data, does email type provide any useful value in classifying email as spam or not spam?

Learning objectives

1. Use a one-way table and a bar chart to summarize a categorical variable. Use counts (frequency) or proportions (relative frequency).
2. Compare distributions of a categorical variable using a two-way table and a side-by-side bar chart, segmented bar chart, or mosaic plot.
3. Calculate marginal and joint frequencies for two-way tables.

2.4.1 Contingency tables and bar charts

Figure 2.33 summarizes two variables: `app_type` and `homeownership`. A table that summarizes data for two categorical variables in this way is called a **contingency table**. Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, the value 3496 corresponds to the number of loans in the data set where the borrower rents their home and the application type was by an individual. Row and column totals are also included. The **row totals** provide the total counts across each row (e.g. $3496 + 3839 + 1170 = 8505$), and **column totals** are total counts down each column. We can also create a table that shows only the overall percentages or proportions for each combination of categories, or we can create a table for a single variable, such as the one shown in Figure 2.34 for the `homeownership` variable.

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

Figure 2.33: A contingency table for `app_type` and `homeownership`.

`homeownership`	Count
rent	3858
mortgage	4789
own	1353
Total	10000

Figure 2.34: A table summarizing the frequencies of each value for the `homeownership` variable.

A **bar chart** (also called bar plot or bar graph) is a common way to display a single categorical variable. The left panel of Figure 2.35 shows a bar chart for the `homeownership` variable. In the right panel, the counts are converted into proportions, showing the proportion of observations that are in each level (e.g. 3858/10000 = 0.3858 for `rent`).

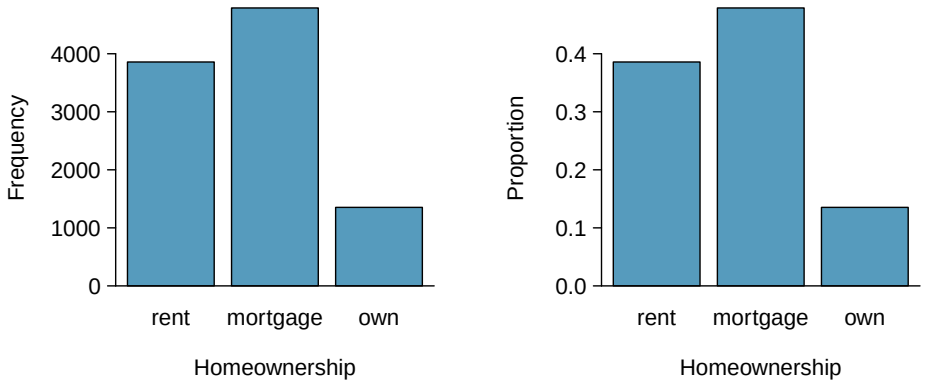


Figure 2.35: Two bar charts of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

2.4.2 Row and column proportions

Sometimes it is useful to understand the fractional breakdown of one variable in another, and we can modify our contingency table to provide such a view. Figure 2.36 shows the **row proportions** for Figure 2.33, which are computed as the counts divided by their row totals. The value 3496 at the intersection of `individual` and `rent` is replaced by 3496/8505 = 0.411, i.e. 3496 divided by its row total, 8505. So what does 0.411 represent? It corresponds to the proportion of individual applicants who rent.

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

Figure 2.36: A contingency table with row proportions for the `app_type` and `homeownership` variables. The row total is off by 0.001 for the `joint` row due to a rounding error.

A contingency table of the column proportions is computed in a similar way, where each **column proportion** is computed as the count divided by the corresponding column total. Figure 2.37 shows such a table, and here the value 0.906 indicates that 90.6% of renters applied as individuals for the loan. This rate is higher compared to loans from people with mortgages (80.2%) or who own their home (85.1%). Because these rates vary between the three levels of `homeownership` (`rent`, `mortgage`, `own`), this provides evidence that the `app_type` and `homeownership` variables are associated.

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

Figure 2.37: A contingency table with column proportions for the `app_type` and `homeownership` variables. The total for the last column is off by 0.001 due to a rounding error.

We could also have checked for an association between `app_type` and `homeownership` in Figure 2.36 using row proportions. When comparing these row proportions, we would look down columns to see if the fraction of loans where the borrower rents, has a mortgage, or owns varied across the `individual` to `joint` application types.

GUIDED PRACTICE 2.73

(a) What does 0.451 represent in Figure 2.36?
(b) What does 0.802 represent in Figure 2.37?[58]

GUIDED PRACTICE 2.74

(a) What does 0.122 at the intersection of `joint` and `own` represent in Figure 2.36?
(b) What does 0.135 represent in the Figure 2.37?[59]

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

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

EXAMPLE 2.75

Data scientists use statistics to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One such characteristic is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is the email format, which indicates whether or not an email has any HTML content, such as bolded text. We'll focus on email format and spam status using the `email` data set, and these variables are summarized in a contingency table in Figure 2.38. Which would be more helpful to someone hoping to classify email as spam or regular email for this table: row or column proportions?

(E)

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

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

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

Figure 2.38: A contingency table for `spam` and `format`.

Example 2.75 points out that row and column proportions are not equivalent. Before settling on one form for a table, it is important to consider each to ensure that the most useful table is constructed. However, sometimes it simply isn't clear which, if either, is more useful.

EXAMPLE 2.76

Look back to Tables 2.36 and 2.37. Are there any obvious scenarios where one might be more useful than the other?

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

2.4.3 Using a bar chart with two variables

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

A **segmented bar chart**, or stacked bar chart, is a graphical display of contingency table information. For example, a segmented bar chart representing Figure 2.37 is shown in Figure 2.39(a), where we have first created a bar chart using the `homeownership` variable and then divided each group by the levels of `app_type`.

One related visualization to the segmented bar chart is the **side-by-side bar chart**, where an example is shown in Figure 2.39(b).

For the last type of bar chart we introduce, the column proportions for the `app_type` and `homeownership` contingency table have been translated into a standardized segmented bar chart in Figure 2.39(c). This type of visualization is helpful in understanding the fraction of individual or joint loan applications for borrowers in each level of `homeownership`. Additionally, since the proportions of `joint` and `individual` vary across the groups, we can conclude that the two variables are associated.

EXAMPLE 2.77

Examine the four bar charts in Figure 2.39. When is the segmented, side-by-side, standardized segmented bar chart, or standardized side-by-side the most useful?

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

Side-by-side bar charts are more agnostic in their display about which variable, if any, represents the explanatory and which the response variable. It is also easy to discern the number of cases in of the six different group combinations. However, one downside is that it tends to require more (E) horizontal space; the narrowness of Figure 2.39(b) makes the plot feel a bit cramped. Additionally, when two groups are of very different sizes, as we see in the `own` group relative to either of the other two groups, it is difficult to discern if there is an association between the variables.

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

The last plot is a standardized side-by-side bar chart. It shows the joint and individual groups as proportions within each level of homeownership, and it offers similar benefits and tradeoffs to the standardized version of the stacked bar plot.

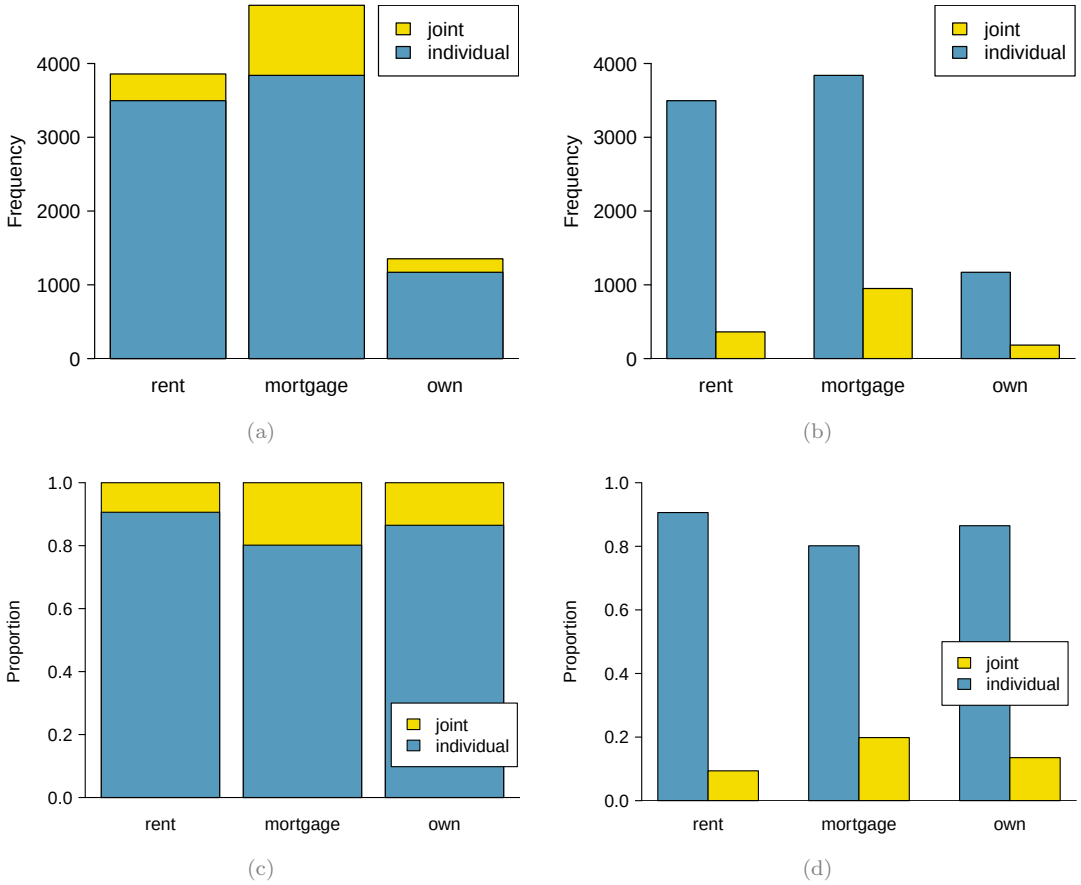


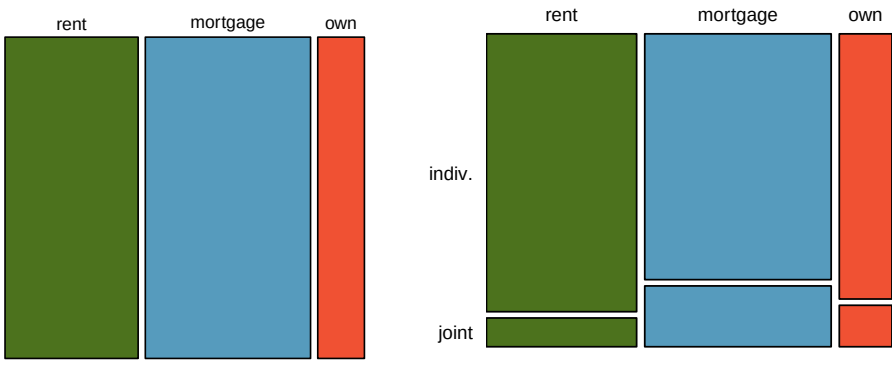
Figure 2.39: (a) segmented bar chart for `homeownership`, where the counts have been further broken down by `app_type`. (b) Side-by-side bar chart for `homeownership` and `app_type`. (c) Standardized version of the segmented bar chart. (d) Standardized side-by-side bar chart. See these bar charts on Tableau Public.

2.4.4 Mosaic plots

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

To get started in creating our first mosaic plot, we'll break a square into columns for each category of the `homeownership` variable, with the result shown in Figure 2.40(a). Each column represents a level of `homeownership`, and the column widths correspond to the proportion of loans in each of those categories. For instance, there are fewer loans where the borrower is an owner than where the borrower has a mortgage. In general, mosaic plots use box *areas* to represent the number of cases in each category.

To create a completed mosaic plot, the single-variable mosaic plot is further divided into pieces in Figure 2.40(b) using the `app_type` variable. Each column is split proportional to the number of loans from individual and joint borrowers. For example, the second column represents loans where the borrower has a mortgage, and it was divided into individual loans (upper) and joint loans (lower). As another example, the bottom segment of the third column represents loans where the borrower owns their home and applied jointly, while the upper segment of this column represents borrowers who are homeowners and filed individually. We can again use this plot to see that the `homeownership` and `app_type` variables are associated, since some columns are divided in different vertical locations than others, which was the same technique used for checking an association in the standardized segmented bar chart.



(a) (b)

Figure 2.40: (a) The one-variable mosaic plot for `homeownership`. (b) Two-variable mosaic plot for both `homeownership` and `app_type`.

In Figure 2.41, we chose to first split by the homeowner status of the borrower. However, we could have instead first split by the application type, as in Figure 2.41. Like with the bar charts, it's common to use the explanatory variable to represent the first split in a mosaic plot, and then for the response to break up each level of the explanatory variable, if these labels are reasonable to attach to the variables under consideration.

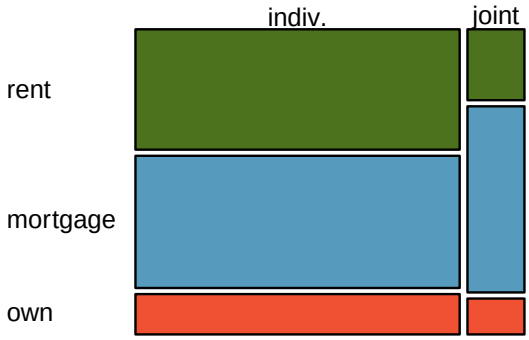


Figure 2.41: Mosaic plot where loans are grouped by the `homeownership` variable after they've been divided into the `individual` and `joint` application types.

2.4.5 The only pie chart you will see in this book

A **pie chart** is shown in Figure 2.42 alongside a bar chart representing the same information. Pie charts can be useful for giving a high-level overview to show how a set of cases break down. However, it is also difficult to decipher details in a pie chart. For example, it takes a couple seconds longer to recognize that there are more loans where the borrower has a mortgage than rent when looking at the pie chart, while this detail is very obvious in the bar chart. While pie charts can be useful, we prefer bar charts for their ease in comparing groups.

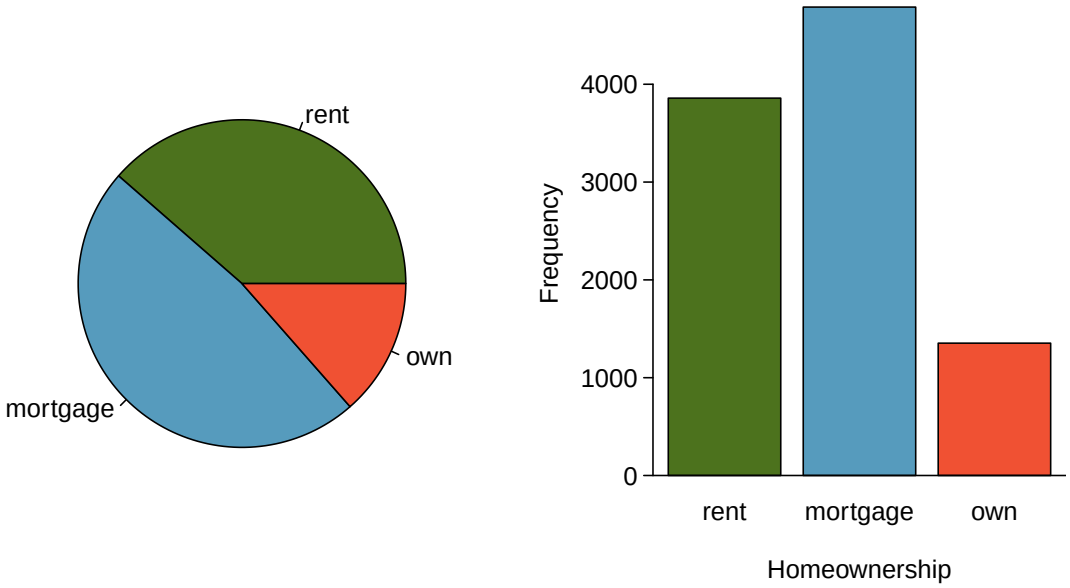


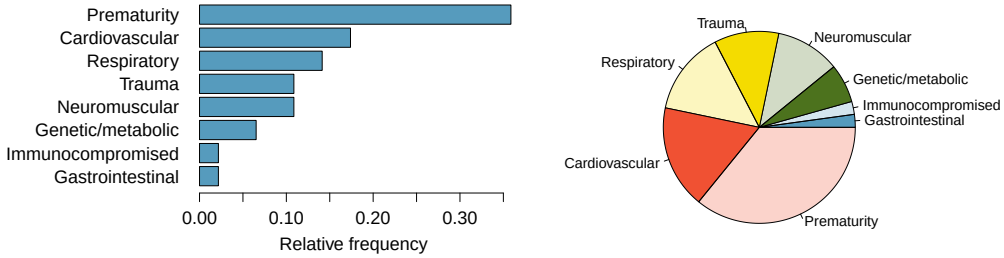
Figure 2.42: A pie chart and bar chart of `homeownership`. Compare multiple ways of summarizing a single categorical variable on Tableau Public.

Section summary

- **Categorical variables**, unlike numerical variables, are simply summarized by **counts** (how many) and **proportions**. These are referred to as frequency and relative frequency, respectively.
- When summarizing one categorical variable, a **one-way frequency table** is useful. For summarizing two categorical variables and their relationship, use a **two-way frequency table** (also known as a contingency table).
- To graphically summarize a single categorical variable, use a **bar chart**. To summarize and compare two categorical variables, use a **side-by-side bar chart**, a **segmented bar chart**, or a **mosaic plot**.
- **Pie charts** are another option for summarizing categorical data, but they are more difficult to read and bar charts are generally a better option.

Exercises

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



(a) What features are apparent in the bar plot but not in the pie chart?

(b) What features are apparent in the pie chart but not in the bar plot?

(c) Which graph would you prefer to use for displaying these categorical data?

2.38 Views on immigration. 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.[60]

		Political ideology			
		Conservative	Moderate	Liberal	Total
Response	(i) Apply for citizenship	57	120	101	278
	(ii) Guest worker	121	113	28	262
	(iii) Leave the country	179	126	45	350
	(iv) Not sure	15	4	1	20
	Total	372	363	175	910

(a) What percent of these Tampa, FL voters identify themselves as conservatives?

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

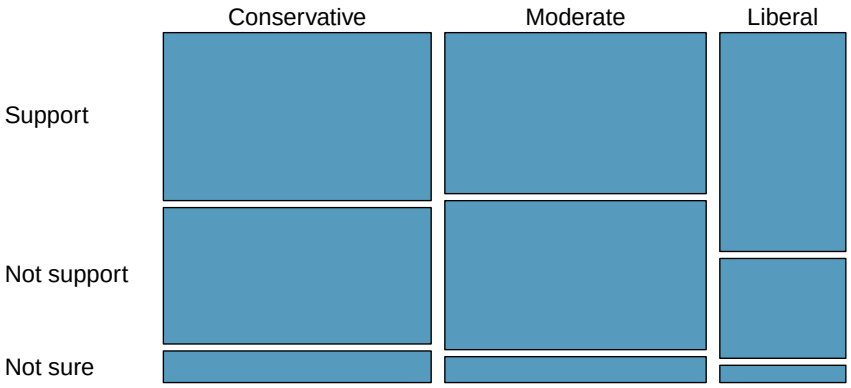
(c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

(d) What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

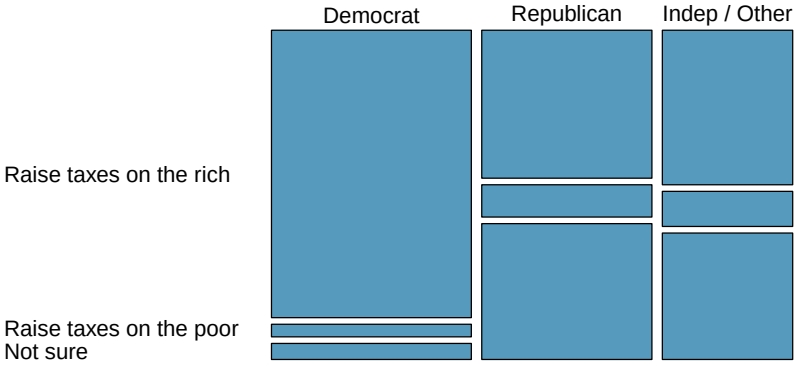
(e) Do political ideology and views on immigration appear to be independent? Explain your reasoning.

[60] **survey:immigFL:2012**.

2.39 Views on the DREAM Act. A random sample of registered voters from Tampa, FL were asked if they support the DREAM Act, a proposed law which would provide a path to citizenship for people brought illegally to the US as children. The survey also collected information on the political ideology of the respondents. Based on the mosaic plot shown below, do views on the DREAM Act and political ideology appear to be independent? Explain your reasoning.[61]



2.40 Raise taxes. A random sample of registered voters nationally were asked whether they think it's better to raise taxes on the rich or raise taxes on the poor. The survey also collected information on the political party affiliation of the respondents. Based on the mosaic plot shown below, do views on raising taxes and political affiliation appear to be independent? Explain your reasoning.[62]



[61] **survey:immigFL:2012**.
[62] **survey:raiseTaxes:2015**.

2.5 Case study: malaria vaccine (special topic)

How large does an observed difference need to be for it to provide convincing evidence that something real is going on, something beyond random variation? Answering this question requires the tools that we will encounter in the later chapters on probability and inference. However, this is such an interesting and important question, and we'll also address it here using simulation. This section can be covered now or in tandem with Chapter **??**: Foundations for Inference.

Learning objectives

1. Recognize that an observed difference in sample statistics may be due to random chance and that we use hypothesis testing to determine if this difference statistically significant (i.e. too large to be attributed to random chance).
2. Set up competing hypotheses and use the results of a simulation to evaluate the degree of support the data provide against the null hypothesis and for the alternative hypothesis.

2.5.1 Variability within data

EXAMPLE 2.78

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

(E)

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

GUIDED PRACTICE 2.79

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?[63]

(G)

We consider a study on a new malaria vaccine called PfSPZ. In this study, volunteer patients were randomized into one of two experiment groups: 14 patients received an experimental vaccine or 6 patients received a placebo vaccine. Nineteen weeks later, all 20 patients were exposed to a drug-sensitive malaria virus strain; the motivation of using a drug-sensitive strain of virus here is for ethical considerations, allowing any infections to be treated effectively. The results are summarized in Figure 2.43, where 9 of the 14 treatment patients remained free of signs of infection while all of the 6 patients in the control group patients showed some baseline signs of infection.

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

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

Figure 2.43: Summary results for the malaria vaccine experiment.

GUIDED PRACTICE 2.80

G Is this an observational study or an experiment? What implications does the study type have on what can be inferred from the results?[64]

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

EXAMPLE 2.81

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

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

Example 2.81 is a reminder that the observed outcomes in the data sample may not perfectly reflect the true relationships between variables since there is **random noise**. While the observed difference in rates of infection is large, the sample size for the study is small, making it unclear if this observed difference represents efficacy of the vaccine or whether it is simply due to chance. We label these two competing claims, H_0 and H_A, which are spoken as "H-nought" and "H-A":

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

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

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

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

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

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

2.5.2 Simulating the study

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

Figure 2.43 shows that 11 patients developed infections and 9 did not. For our simulation, we will suppose the infections were independent of the vaccine and we were able to *rewind* back to when the researchers randomized the patients in the study. If we happened to randomize the patients differently, we may get a different result in this hypothetical world where the vaccine doesn't influence the infection. Let's complete another **randomization** using a simulation.

In this **simulation**, we take 20 notecards to represent the 20 patients, where we write down "infection" on 11 cards and "no infection" on 9 cards. In this hypothetical world, we believe each patient that got an infection was going to get it regardless of which group they were in, so let's see what happens if we randomly assign the patients to the treatment and control groups again. We thoroughly shuffle the notecards and deal 14 into a `vaccine` pile and 6 into a `placebo` pile. Finally, we tabulate the results, which are shown in Figure 2.44.

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

Figure 2.44: Simulation results, where any difference in infection rates is purely due to chance.

GUIDED PRACTICE 2.82

G What is the difference in infection rates between the two simulated groups in Figure 2.44? How does this compare to the observed 64.3% difference in the actual data?[65]

2.5.3 Checking for independence

We computed one possible difference under the independence model in Guided Practice 2.82, which represents one difference due to chance. While in this first simulation, we physically dealt out notecards to represent the patients, it is more efficient to perform this simulation using a computer. Repeating the simulation on a computer, we get another difference due to chance:

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

And another:

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

And so on until we repeat the simulation enough times that we have a good idea of what represents the *distribution of differences from chance alone*. Figure 2.45 shows a stacked plot of the differences found from 100 simulations, where each dot represents a simulated difference between the infection rates (control rate minus treatment rate).

Note that the distribution of these simulated differences is centered around 0. We simulated these differences assuming that the independence model was true, and under this condition, we

[65] $4/6 - 7/14 = 0.167$ or about 16.7% in favor of the vaccine. This difference due to chance is much smaller than the difference observed in the actual groups.

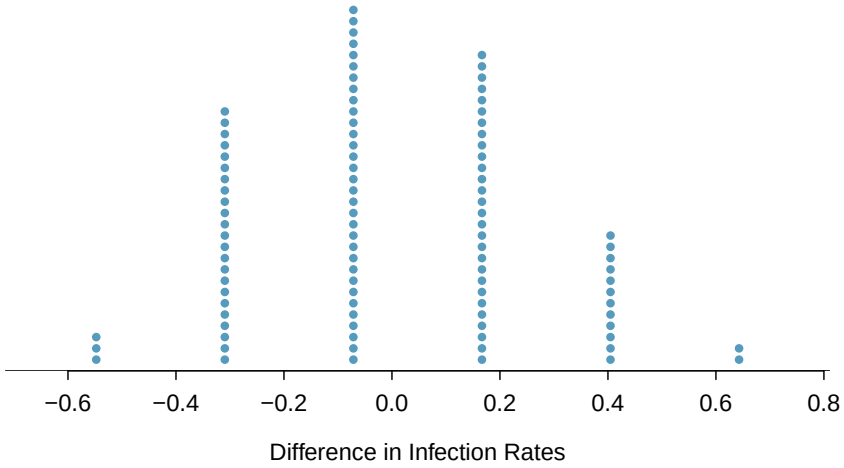


Figure 2.45: A stacked dot plot of differences from 100 simulations produced under the independence model, H_0, where in these simulations infections are unaffected by the vaccine. Two of the 100 simulations had a difference of at least 64.3%, the difference observed in the study.

expect the difference to be near zero with some random fluctuation, where *near* is pretty generous in this case since the sample sizes are so small in this study.

EXAMPLE 2.83

Given the results of the simulation shown in Figure 2.45, about how often would you expect to observe a result as large as 64.3% if H_0 were true?

(E)

Because a result this large happened 2 times out the 100 simulations, we would expect such a large value only 2% of the time if H_0 were true.

There are two possible interpretations of the results of the study:

H_0 **Independence model.** The vaccine has no effect on infection rate, and we just happened to observe a rare event.

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

Based on the simulations, we have two options. (1) We conclude that the study results do not provide strong enough evidence against the independence model, meaning we do not conclude that the vaccine had an effect in this clinical setting. (2) We conclude the evidence is sufficiently strong to reject H_0, and we assert that the vaccine was useful.

Is 2% small enough to make us reject the independence model? That depends on how much evidence we require. The smaller that probability is, the more evidence it provides against H_0. Later, we will see that researchers often use a cutoff of 5%, though it can depend upon the situation. Using the 5% cutoff, we would reject the independence model in favor of the alternative. That is, we are concluding the data provide strong evidence that the vaccine provides some protection against malaria in this clinical setting.

When there is strong enough evidence that the result points to a real difference and is not simply due to random variation, we call the result **statistically significant**.

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

Exercises

2.41 Side effects of Avandia. Rosiglitazone is the active ingredient in the controversial type 2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is pioglitazone, the active ingredient in a diabetes medicine called Actos. In a nationwide retrospective observational study of 227,571 Medicare beneficiaries aged 65 years or older, it was found that 2,593 of the 67,593 patients using rosiglitazone and 5,386 of the 159,978 using pioglitazone had serious cardiovascular problems. These data are summarized in the contingency table below.[66]

		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

(a) Determine if each of the following statements is true or false. If false, explain why. *Be careful:* The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.

i. Since more patients on pioglitazone had cardiovascular problems (5,386 vs. 2,593), we can conclude that the rate of cardiovascular problems for those on a pioglitazone treatment is higher.

ii. The data suggest that diabetic patients who are taking rosiglitazone are more likely to have cardiovascular problems since the rate of incidence was (2,593 / 67,593 = 0.038) 3.8% for patients on this treatment, while it was only (5,386 / 159,978 = 0.034) 3.4% for patients on pioglitazone.

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

iv. Based on the information provided so far, we cannot tell if the difference between the rates of incidences is due to a relationship between the two variables or due to chance.

(b) What proportion of all patients had cardiovascular problems?

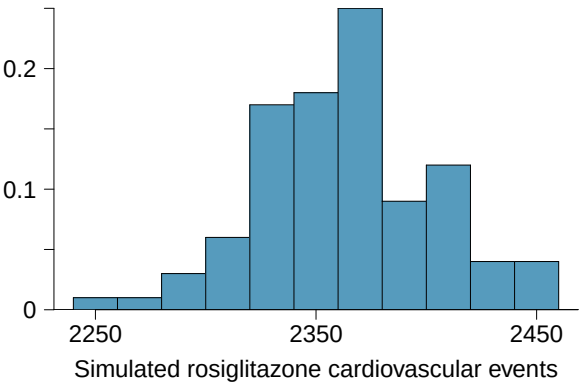
(c) If the type of treatment and having cardiovascular problems were independent, about how many patients in the rosiglitazone group would we expect to have had cardiovascular problems?

(d) We can investigate the relationship between outcome and treatment in this study using a randomization technique. While in reality we would carry out the simulations required for randomization using statistical software, suppose we actually simulate using index cards. In order to simulate from the independence model, which states that the outcomes were independent of the treatment, we write whether or not each patient had a cardiovascular problem on cards, shuffled all the cards together, then deal them into two groups of size 67,593 and 159,978. We repeat this simulation 1,000 times and each time record the number of people in the rosiglitazone group who had cardiovascular problems. Use the relative frequency histogram of these counts to answer (i)-(iii).

i. What are the claims being tested?

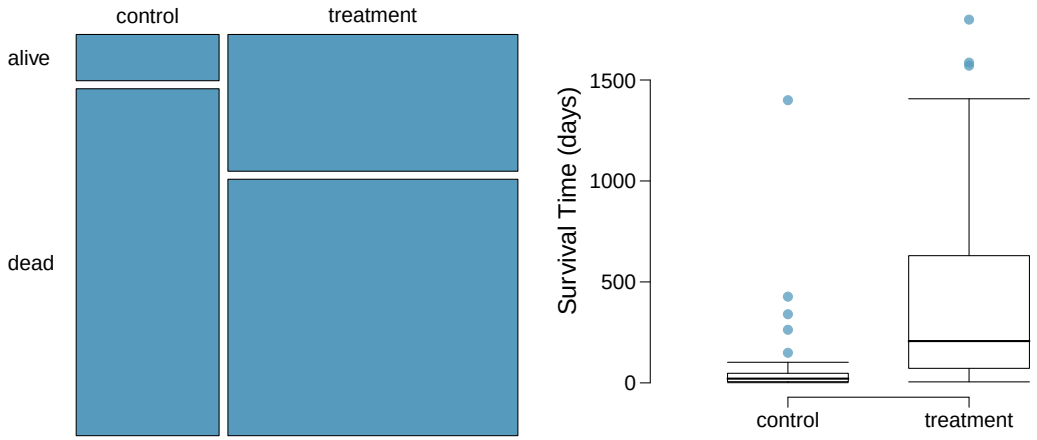
ii. Compared to the number calculated in part (b), which would provide more support for the alternative hypothesis, *more* or *fewer* patients with cardiovascular problems in the rosiglitazone group?

iii. What do the simulation results suggest about the relationship between taking rosiglitazone and having cardiovascular problems in diabetic patients?



[66] **Graham:2010**.

2.42 Heart transplants. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable `transplant` indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died. Another variable called `survived` was used to indicate whether or not the patient was alive at the end of the study.[67]



(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.

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

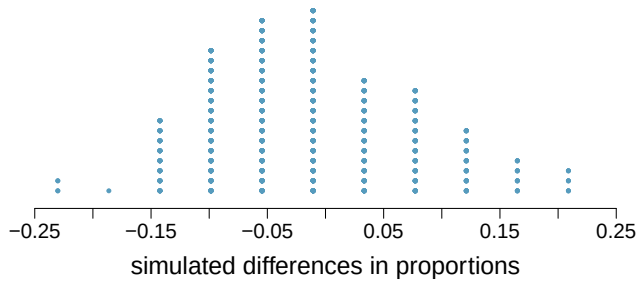
(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested?

ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

> We write *alive* on _____ cards representing patients who were alive at the end of the study, and *dead* on _____ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size _____ representing treatment, and another group of size _____ representing control. We calculate the difference between the proportion of *dead* cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at _____. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are _____. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?



[67] **Turnbull+Brown+Hu:1974**.

Chapter highlights

A raw data matrix/table may have thousands of rows. The data need to be summarized in order to makes sense of all the information. In this chapter, we looked at ways to summarize data **graphically**, **numerically**, and **verbally**.

Categorical data

- A single **categorical variable** is summarized with **counts** or **proportions**proportion in a **one-way table**. A **bar chart** is used to show the frequency or relative frequency of the categories that the variable takes on.
- Two categorical variables can be summarized in a **two-way table** and with a **side-by-side bar chart** or a **segmented bar chart**.

Numerical data

- When looking at a single **numerical variable**, we try to understand the **distribution** of the variable. The distribution of a variable can be represented with a frequency table and with a graph, such as a **stem-and-leaf plot** or **dot plot** for small data sets, or a **histogram** for larger data sets. If only a summary is desired, a **box plot** may be used.
- The **distribution** of a variable can be described and summarized with **center** (mean or median), **spread** (SD or IQR), and **shape** (right skewed, left skewed, approximately symmetric).
- **Z-scores** and **percentiles** are useful for identifying a data point's relative position within a data set.
- When a distribution is nearly normal, we can use the **empirical rule** (68-95-99.7 rule) and we can use **normal approximation** to approximate area/percent under a histogram using area/percent under the normal curve.
- **Outliers** are values that appear extreme relative to the rest of the data. Investigating outliers can provide insight into properties of the data or may reveal data collection/entry errors.
- When **comparing the distribution** of two variables, use two dot plots, two histograms, a back-to-back stem-and-leaf, or parallel box plots.
- To look at the **association** between two numerical variables, use a **scatterplot**.

Graphs and numbers can summarize data, but they alone are insufficient. It is the role of the researcher or data scientist to ask questions, to use these tools to identify patterns and departure from patterns, and to make sense of this in the context of the data. Strong writing skills are critical for being able to communicate the results to a wider audience.

Chapter exercises

2.43 Make-up exam. In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found an average score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.

(a) Does the new student's score increase or decrease the average score?

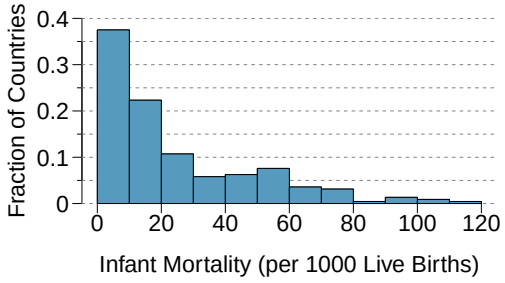
(b) What is the new average?

(c) Does the new student's score increase or decrease the standard deviation of the scores?

2.44 Infant mortality. The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014.[68]

(a) Estimate Q1, the median, and Q3 from the histogram.

(b) Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.



2.45 TV watchers. Students in an AP Statistics class were asked how many hours of television they watch per week (including online streaming). This sample yielded an average of 4.71 hours, with a standard deviation of 4.18 hours. Is the distribution of number of hours students watch television weekly symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

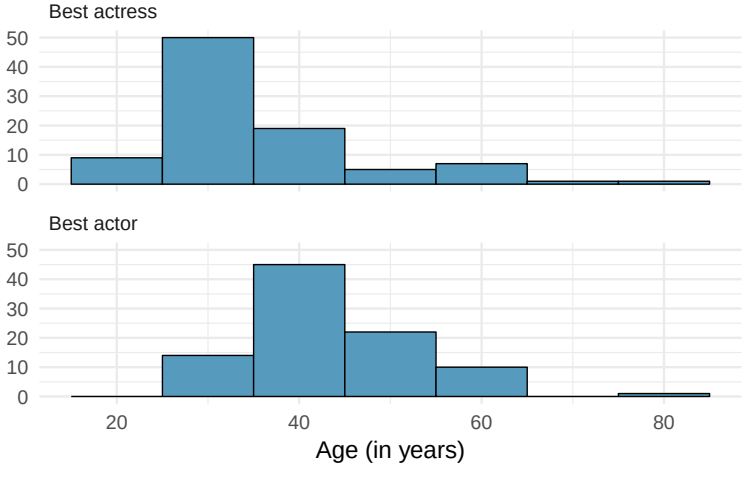
2.46 A new statistic. The statistic $\frac{\bar{x}}{median}$ can be used as a measure of skewness. Suppose we have a distribution where all observations are greater than 0, $x_i > 0$. What is the expected shape of the distribution under the following conditions? Explain your reasoning.

(a) $\frac{\bar{x}}{median} = 1$

(b) $\frac{\bar{x}}{median} < 1$

(c) $\frac{\bar{x}}{median} > 1$

2.47 Oscar winners. The first Oscar awards for best actor and best actress were given out in 1929. The histograms below show the age distribution for all of the best actor and best actress winners from 1929 to 2018. Summary statistics for these distributions are also provided. Compare the distributions of ages of best actor and actress winners.[69]



	Best Actress
Mean	36.2
SD	11.9
n	92

	Best Actor
Mean	43.8
SD	8.83
n	92

[68] **data:ciaFactbook**.
[69] **data:oscars**.

2.48 Exam scores. The average on a history exam (scored out of 100 points) was 85, with a standard deviation of 15. Is the distribution of the scores on this exam symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

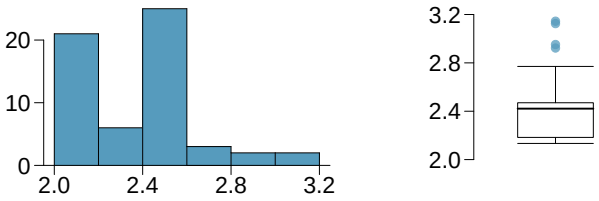
2.49 Stats scores. Below are the final exam scores of twenty introductory statistics students.

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

Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

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

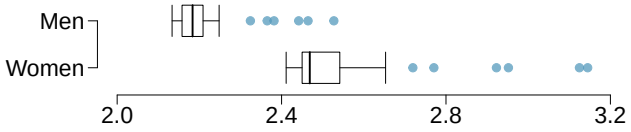
2.50 Marathon winners. The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.



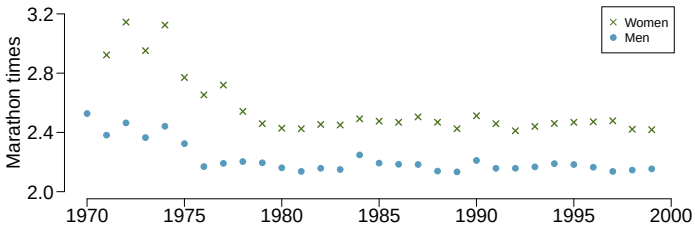
(a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

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

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



(d) The time series plot shown below is another way to look at these data. Describe what is visible in this plot but not in the others.



2.51 Birth weight. In a large study of birth weight of newborns, the weights of 23,419 newborn boys were recorded. The distribution of weights was approximately normal with a mean of 7.44 lbs (3376 grams) and a standard deviation of 1.33 lbs (603 grams). The government classifies a newborn as having low birth weight if the weight is less than 5.5 pounds.[70]

(a) What percent of these newborns had a low birth weight?

(b) Approximately what percent of these babies weighed greater than 10 pounds?

(c) Approximately *how many* of these newborns weighed greater than 10 pounds?

(d) How much would a newborn have to weigh in order to be at the 90th percentile among this group?

[70] www.biomedcentral.com/1471-2393/8/5

Chapter 3

Probability and probability distributions

3.1 Defining probability

3.2 Conditional probability

3.3 Simulations

3.4 Random variables

3.5 Geometric distribution

3.6 Binomial distribution

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

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

3.1 Defining probability

What is the probability of rolling an even number on a die? Of getting 5 heads in row when tossing a coin? Of drawing a Heart or an Ace from a deck of cards? The study of probability is fun and interesting in its own right, but it also forms the foundation for statistical models and inferential procedures, many of which we will investigate in subsequent chapters.

Learning objectives

1. Describe the long-run relative frequency interpretation of probability and understand its relationship to the "Law of Large Numbers".
2. Use Venn diagrams to represent events and their probabilities and to visualize the complement, union, and intersection of events.
3. Use the General Addition Rule to find the probability that at least one of several events occurs.
4. Understand when events are disjoint (mutually exclusive) and how that simplifies the General Addition Rule.
5. Apply the Multiplication Rule for finding the joint probability of independent events.

3.1.1 Introductory examples

EXAMPLE 3.1

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

(E)

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

EXAMPLE 3.2

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

(E)

---

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

EXAMPLE 3.3

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

(E)

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

EXAMPLE 3.4

What is the chance of not rolling a 2?

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

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

EXAMPLE 3.5

Consider rolling two dice. If $1/6^{th}$ 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 `1`s?

(E)

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

3.1.2 Probability

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

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

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

PROBABILITY

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

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

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

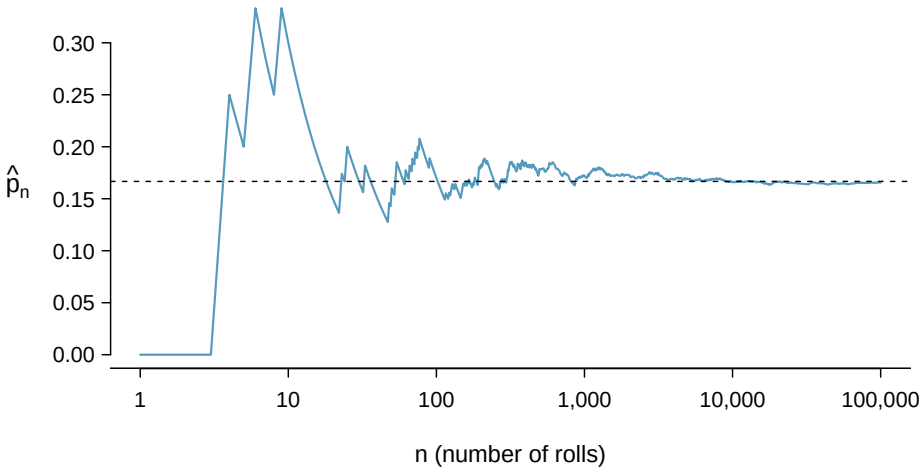


Figure 3.1: The fraction of die rolls that are 1 at each stage in a simulation. The relative frequency tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.

LAW OF LARGE NUMBERS

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

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

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

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

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

GUIDED PRACTICE 3.6

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

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

MODELING A PROCESS AS RANDOM

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

3.1.3 Disjoint or mutually exclusive outcomes

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

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

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

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.

[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`.

ADDITION RULE OF DISJOINT OUTCOMES

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

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

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

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

GUIDED PRACTICE 3.7

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

GUIDED PRACTICE 3.8

G In the `email` data set in Chapter 2, the `number` variable described whether no number (labeled `none`), only one or more small numbers (`small`), or whether at least one big number appeared in an email (`big`). Of the 3,921 emails, 549 had no numbers, 2,827 had only one or more small numbers, and 545 had at least one big number. (a) Are the outcomes `none`, `small`, and `big` disjoint? (b) Determine the proportion of emails with value `small` and `big` separately. (c) Use the Addition Rule for disjoint outcomes to compute the probability a randomly selected email from the data set has a number in it, small or big.[3]

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

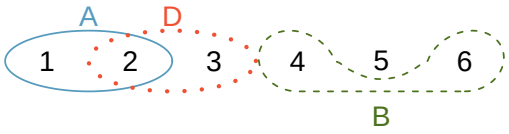


Figure 3.2: Three events, A, B, and D, consist of outcomes from rolling a die. A and B are disjoint since they do not have any outcomes in common.

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

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

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

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

GUIDED PRACTICE 3.9

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

GUIDED PRACTICE 3.10

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

GUIDED PRACTICE 3.11

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

3.1.4 Probabilities when events are not disjoint

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

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

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

GUIDED PRACTICE 3.12

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

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

[4](a) $P(A) = P(\texttt{1 or 2}) = P(\texttt{1}) + P(\texttt{2}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$. (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 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 $P(\text{face card}) = \frac{12}{52} = \frac{3}{13} = 0.231$.

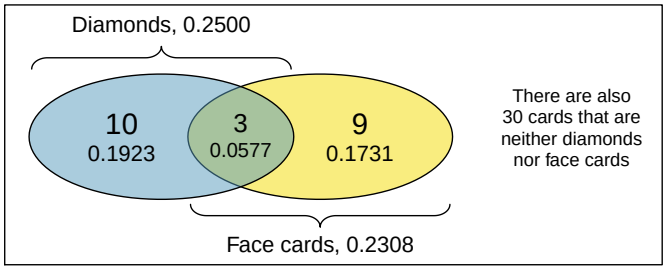


Figure 3.4: A Venn diagram for diamonds and face cards.

GUIDED PRACTICE 3.13

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

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

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

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

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

Equation (3.14) is an example of the **General Addition Rule**.

GENERAL ADDITION RULE

If A and B are any two events, disjoint or not, then the probability that A or B will occur is

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

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

SYMBOLIC NOTATION FOR "AND" AND "OR"

The symbol $\cap$ means intersection and is equivalent to "and".
The symbol $\cup$ means union and is equivalent to "or".
It is common to see the General Addition Rule written as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

"OR" IS INCLUSIVE

When we write, "or" in statistics, we mean "and/or" unless we explicitly state otherwise. Thus, A or B occurs means A, B, or both A and B occur. This is equivalent to at least one of A or B occurring.

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

GUIDED PRACTICE 3.14

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

GUIDED PRACTICE 3.15

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

GUIDED PRACTICE 3.16

(a) Use your Venn diagram from Guided Practice 3.15 to determine the probability a randomly drawn email from the `email` data set is spam and had small numbers (but not big numbers). (b) What is the probability that the email had either of these attributes?[12]

3.1.5 Complement of an event

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

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is `2` or `3`. Then the **complement** represents all outcomes in our sample space that are not in D, which is denoted by $D^c = \{1, 4, 5, 6\}$. That is, D^c is the set of all possible outcomes not already included in D. Figure 3.5 shows the relationship between D, D^c, and the sample space S.

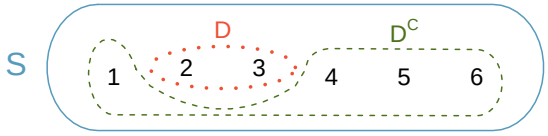


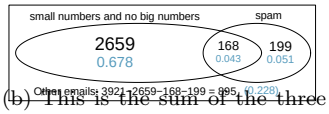
Figure 3.5: Event $D = \{2, 3\}$ and its complement, $D^c = \{1, 4, 5, 6\}$. S represents the sample space, which is the set of all possible events.

GUIDED PRACTICE 3.17

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

[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 (3.14) is 0 (see part (a)) and we are left with the Addition Rule for disjoint events.

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



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

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

GUIDED PRACTICE 3.18

Events $A = \{1, 2\}$ and $B = \{4, 6\}$ are shown in Figure 3.2 on page 88. (a) Write out what A^c and B^c represent. (b) Compute $P(A^c)$ and $P(B^c)$. (c) Compute $P(A) + P(A^c)$ and $P(B) + P(B^c)$.[14]

An event A together with its complement A^c comprise the entire sample space. Because of this we can say that $P(A) + P(A^c) = 1$.

COMPLEMENT

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

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

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

GUIDED PRACTICE 3.19

A die is rolled 10 times. (a) What is the complement of getting at least one 6 in 10 rolls of the die? (b) What is the complement of getting at most three 6's in 10 rolls of the die?[15]

3.1.6 Independence

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

Example 3.5 provides a basic example of two independent processes: rolling two dice. We want to determine the probability that both will be 1. Suppose one of the dice is red and the other white. If the outcome of the red die is a 1, it provides no information about the outcome of the white die. We first encountered this same question in Example 3.5 (page 86), where we calculated the probability using the following reasoning: $1/6^{th}$ of the time the red die is a 1, and $1/6^{th}$ of *those* times the white die will also be 1. This is illustrated in Figure 3.6. Because the rolls are independent, the probabilities of the corresponding outcomes can be multiplied to get the final answer: $(1/6) \times (1/6) = 1/36$. This can be generalized to many independent processes.

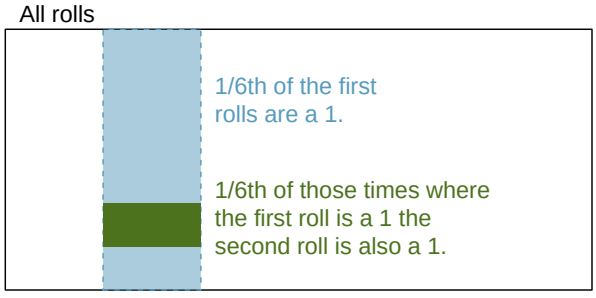


Figure 3.6: $1/6^{th}$ of the time, the first roll is a 1. Then $1/6^{th}$ of *those* times, the second roll will also be a 1.

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

[15] (a) The complement of getting at least one 6 in ten rolls of a die is getting zero 6's in the 10 rolls. (b) The complement of getting at most three 6's in 10 rolls is getting four, five, ..., nine, or ten 6's in 10 rolls.

EXAMPLE 3.20

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

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

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

Examples 3.5 and 3.20 illustrate what is called the Multiplication Rule for independent processes.

MULTIPLICATION RULE FOR INDEPENDENT PROCESSES

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

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

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

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

GUIDED PRACTICE 3.21

(G) About 9% of people are left-handed. Suppose 2 people are selected at random from the U.S. population. Because the sample size of 2 is very small relative to the population, it is reasonable to assume these two people are independent. (a) What is the probability that both are left-handed? (b) What is the probability that both are right-handed?[16]

GUIDED PRACTICE 3.22

Suppose 5 people are selected at random.[17]

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

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

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

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

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

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

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

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

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

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

Suppose the variables `handedness` and `gender` are independent, i.e. knowing someone's `gender` provides no useful information about their `handedness` and vice-versa. Then we can compute whether a randomly selected person is right-handed and female[18] using the Multiplication Rule:

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

GUIDED PRACTICE 3.23

Three people are selected at random.[19]

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

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

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

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

Sometimes we wonder if one outcome provides useful information about another outcome. The question we are asking is, are the occurrences of the two events independent? We say that two events A and B are independent if they satisfy Equation (3.21).

EXAMPLE 3.24

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 3.21 is satisfied:

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

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

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

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

Section summary

- When an outcome depends upon a chance process, we can define the **probability** of the outcome as the proportion of times it would occur if we repeated the process an infinite number of times. Also, even when an outcome is not truly random, modeling it with probability can be useful.
- The **Law of Large Numbers** states that the **relative frequency**, or proportion of times an outcome occurs after n repetitions, stabilizes around the true probability as n gets large.
- The probability of an event is always between 0 and 1, inclusive.
- The probability of an event and the probability of its **complement** add up to 1. Sometime we use $P(A) = 1 - P(\text{not } A)$ when $P(\text{not } A)$ is easier to calculate than $P(A)$.
- A and B are **disjoint**, i.e. **mutually exclusive**, if they cannot happen together. In this case, the events do not overlap and $P(A \text{ and } B) = 0$.
- In the *special case* where A and B are **disjoint** events: $P(A \text{ or } B) = P(A) + P(B)$.
- When A and B are not disjoint, adding $P(A)$ and $P(B)$ will overestimate $P(A \text{ or } B)$ because the overlap of A and B will be added twice. Therefore, when A and B are not disjoint, use the **General Addition Rule**:
  $P(A \text{ or B}) = P(A) + P(B) - P(A \text{ and } B)$.[20]
- To find the probability that *at least one* of several events occurs, use a special case of the rule of **complements**: $P(\text{at least one}) = 1 - P(\text{none})$.
- When only considering two events, the probability that one *or* the other happens is equal to the probability that *at least one* of the two events happens. When dealing with more than two events, the General Addition Rule becomes very complicated. Instead, to find the probability that A or B or C occurs, find the probability that none of them occur and subtract that value from 1.
- Two events are **independent** when the occurrence of one does not change the likelihood of the other.
- In the *special case* where A and B are **independent**: $P(A \text{ and } B) = P(A) \times P(B)$.

[20] Often written: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Exercises

3.1 True or false. Determine if the statements below are true or false, and explain your reasoning.

(a) If a fair coin is tossed many times and the last eight tosses are all heads, then the chance that the next toss will be heads is somewhat less than 50%.

(b) Drawing a face card (jack, queen, or king) and drawing a red card from a full deck of playing cards are mutually exclusive events.

(c) Drawing a face card and drawing an ace from a full deck of playing cards are mutually exclusive events.

3.2 Roulette wheel. The game of roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball.

(a) You watch a roulette wheel spin 3 consecutive times and the ball lands on a red slot each time. What is the probability that the ball will land on a red slot on the next spin?

(b) You watch a roulette wheel spin 300 consecutive times and the ball lands on a red slot each time. What is the probability that the ball will land on a red slot on the next spin?

(c) Are you equally confident of your answers to parts (a) and (b)? Why or why not?

Photo by Håkan Dahlström (http://flic.kr/p/93fEzp) CC BY 2.0 license

3.3 Four games, one winner. Below are four versions of the same game. Your archnemesis gets to pick the version of the game, and then you get to choose how many times to flip a coin: 10 times or 100 times. Identify how many coin flips you should choose for each version of the game. It costs $1 to play each game. Explain your reasoning.

(a) If the proportion of heads is larger than 0.60, you win $1.

(b) If the proportion of heads is larger than 0.40, you win $1.

(c) If the proportion of heads is between 0.40 and 0.60, you win $1.

(d) If the proportion of heads is smaller than 0.30, you win $1.

3.4 Backgammon. Backgammon is a board game for two players in which the playing pieces are moved according to the roll of two dice. Players win by removing all of their pieces from the board, so it is usually good to roll high numbers. You are playing backgammon with a friend and you roll two 6s in your first roll and two 6s in your second roll. Your friend rolls two 3s in his first roll and again in his second row. Your friend claims that you are cheating, because rolling double 6s twice in a row is very unlikely. Using probability, show that your rolls were just as likely as his.

3.5 Coin flips. If you flip a fair coin 10 times, what is the probability of

(a) getting all tails?

(b) getting all heads?

(c) getting at least one tails?

3.6 Dice rolls. If you roll a pair of fair dice, what is the probability of

(a) getting a sum of 1?

(b) getting a sum of 5?

(c) getting a sum of 12?

3.7 Swing voters. A Pew Research survey asked 2,373 randomly sampled registered voters their political affiliation (Republican, Democrat, or Independent) and whether or not they identify as swing voters. 35% of respondents identified as Independent, 23% identified as swing voters, and 11% identified as both.[21]

(a) Are being Independent and being a swing voter disjoint, i.e. mutually exclusive?

(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

(c) What percent of voters are Independent but not swing voters?

(d) What percent of voters are Independent or swing voters?

(e) What percent of voters are neither Independent nor swing voters?

(f) Is the event that someone is a swing voter independent of the event that someone is a political Independent?

3.8 Poverty and language. The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 20.7% speak a language other than English (foreign language) at home, and 4.2% fall into both categories.[22]

(a) Are living below the poverty line and speaking a foreign language at home disjoint?

(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

(c) What percent of Americans live below the poverty line and only speak English at home?

(d) What percent of Americans live below the poverty line or speak a foreign language at home?

(e) What percent of Americans live above the poverty line and only speak English at home?

(f) Is the event that someone lives below the poverty line independent of the event that the person speaks a foreign language at home?

3.9 Disjoint vs. independent. In parts (a) and (b), identify whether the events are disjoint, independent, or neither (events cannot be both disjoint and independent).

(a) You and a randomly selected student from your class both earn A's in this course.

(b) You and your class study partner both earn A's in this course.

(c) If two events can occur at the same time, must they be dependent?

3.10 Guessing on an exam. In a multiple choice exam, there are 5 questions and 4 choices for each question (a, b, c, d). Nancy has not studied for the exam at all and decides to randomly guess the answers. What is the probability that:

(a) the first question she gets right is the 5^{th} question?

(b) she gets all of the questions right?

(c) she gets at least one question right?

[21] **indepSwing**.
[22] **poorLang**.

3.11 Educational attainment of couples. The table below shows the distribution of education level attained by US residents by gender based on data collected in the 2010 American Community Survey.[23]

		Gender	
		Male	Female
Highest education attained	Less than 9th grade	0.07	0.13
	9th to 12th grade, no diploma	0.10	0.09
	HS graduate (or equivalent)	0.30	0.20
	Some college, no degree	0.22	0.24
	Associate's degree	0.06	0.08
	Bachelor's degree	0.16	0.17
	Graduate or professional degree	0.09	0.09
	Total	1.00	1.00

(a) What is the probability that a randomly chosen man has at least a Bachelor's degree?

(b) What is the probability that a randomly chosen woman has at least a Bachelor's degree?

(c) What is the probability that a man and a woman getting married both have at least a Bachelor's degree? Note any assumptions you must make to answer this question.

(d) If you made an assumption in part (c), do you think it was reasonable? If you didn't make an assumption, double check your earlier answer and then return to this part.

3.12 School absences. Data collected at elementary schools in DeKalb County, GA suggest that each year roughly 25% of students miss exactly one day of school, 15% miss 2 days, and 28% miss 3 or more days due to sickness.[24]

(a) What is the probability that a student chosen at random doesn't miss any days of school due to sickness this year?

(b) What is the probability that a student chosen at random misses no more than one day?

(c) What is the probability that a student chosen at random misses at least one day?

(d) If a parent has two kids at a DeKalb County elementary school, what is the probability that neither kid will miss any school? Note any assumption you must make to answer this question.

(e) If a parent has two kids at a DeKalb County elementary school, what is the probability that both kids will miss some school, i.e. at least one day? Note any assumption you make.

(f) If you made an assumption in part (d) or (e), do you think it was reasonable? If you didn't make any assumptions, double check your earlier answers.

[23] **eduSex**.
[24] **Mizan:2011**.

3.2 Conditional probability

In this section we will use conditional probabilities to answer the following questions:

- What is the likelihood that a machine learning algorithm will misclassify a photo as being about fashion if it is not actually about fashion?
- How much more likely are children to attend college whose parents attended college than children whose parents did not attend college?
- Given that a person receives a positive test result for a disease, what is the probability that the person actually has the disease?

Learning objectives

1. Understand conditional probability and how to calculate it.
2. Calculate joint and conditional probabilities based on a two-way table.
3. Use the General Multiplication Rule to find the probability of joint events.
4. Determine whether two events are independent and whether they are mutually exclusive, based on the definitions of those terms.
5. Draw a tree diagram with at least two branches to organize possible outcomes and their probabilities. Understand that the second branch represents conditional probabilities.
6. Use the tree diagram or Bayes' Theorem to solve "inverted" conditional probabilities.

3.2.1 Exploring probabilities with a contingency table

The `photo_classify` data set represents a sample of 1822 photos from a photo sharing website. Data scientists have been working to improve a classifier for whether the photo is about fashion or not, and these 659 photos represent a test for their classifier. Each photo gets two classifications: the first is called `mach_learn` and gives a classification from a machine learning (ML) system of either `pred_fashion` or `pred_not`. Each of these 1822 photos have also been classified carefully by a team of people, which we take to be the source of truth; this variable is called `truth` and takes values `fashion` and `not`. Figure 3.7 summarizes the results.

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

Figure 3.7: Contingency table summarizing the `photo_classify` data set.

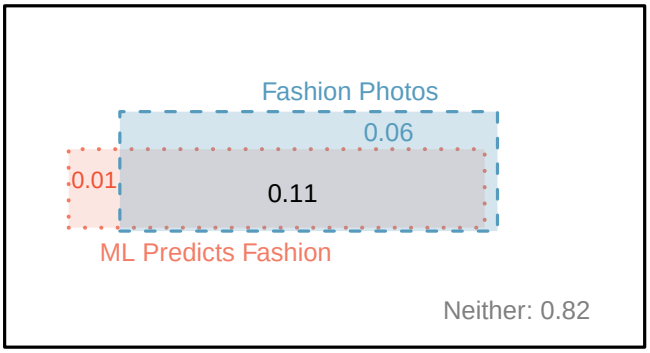


Figure 3.8: A Venn diagram using boxes for the `photo_classify` data set.

EXAMPLE 3.25

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

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

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

EXAMPLE 3.26

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

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

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

3.2.2 Marginal and joint probabilities

Figure 3.7 includes row and column totals for each variable separately in the `photo_classify` data set. These totals represent **marginal probabilities** for the sample, which are the probabilities based on a single variable without regard to any other variables. For instance, a probability based solely on the `mach_learn` variable is a marginal probability:

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

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

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

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

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

means the same thing as

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

MARGINAL AND JOINT PROBABILITIES

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

We use **table proportions** to summarize joint probabilities for the `photo_classify` sample. These proportions are computed by dividing each count in Figure 3.7 by the table's total, 1822, to obtain the proportions in Figure 3.9. The joint probability distribution of the `mach_learn` and `truth` variables is shown in Figure 3.10.

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

Figure 3.9: Probability table summarizing the `photo_classify` data set.

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

Figure 3.10: Joint probability distribution for the `photo_classify` data set.

GUIDED PRACTICE 3.27

G Verify Figure 3.10 represents a probability distribution: events are disjoint, all probabilities are non-negative, and the probabilities sum to 1.[25]

[25] Each of the four outcome combination are disjoint, all probabilities are indeed non-negative, and the sum of the probabilities is $0.1081 + 0.0121 + 0.0615 + 0.8183 = 1.00$.

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

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

3.2.3 Defining conditional probability

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

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

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

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

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

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

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

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

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

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

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

We considered only those cases that satisfied the condition, where the ML algorithm predicted fashion. Of these cases, the conditional probability was the fraction representing the outcome of interest, that the photo was about fashion. Suppose we were provided only the information in Figure 3.9, i.e. only probability data. Then if we took a sample of 1000 photos, we would anticipate about 12.0% or $0.120 \times 1000 = 120$ would be predicted to be about fashion (`mach_learn` is `pred_fashion`). Similarly, we would expect about 10.8% or $0.108 \times 1000 = 108$ to meet both the in-

formation criteria and represent our outcome of interest. Then the conditional probability can be computed as

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

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

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

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

CONDITIONAL PROBABILITY

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

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

GUIDED PRACTICE 3.28

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

(b) Determine the probability from part (a). Figure 3.9 on page 101 may be helpful.[26]

GUIDED PRACTICE 3.29

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

(b) Using the answers from part (a) and Guided Practice 3.28(b), compute

$$\begin{aligned} &P(\texttt{mach\_learn} \text{ is } \texttt{pred\_fashion} \mid \texttt{truth} \text{ is } \texttt{fashion}) \\ &\quad + \ P(\texttt{mach\_learn} \text{ is } \texttt{pred\_not} \mid \texttt{truth} \text{ is } \texttt{fashion}) \end{aligned}$$

(c) Provide an intuitive argument to explain why the sum in (b) is 1.[27]

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

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

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

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

3.2.4 Smallpox in Boston, 1721

The `smallpox` data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston.[28] Doctors at the time believed that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death.

Each case represents one person with two variables: `inoculated` and `result`. The variable `inoculated` takes two levels: `yes` or `no`, indicating whether the person was inoculated or not. The variable `result` has outcomes `lived` or `died`. These data are summarized in Tables 3.11 and 3.12.

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

Figure 3.11: Contingency table for the `smallpox` data set.

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

Figure 3.12: Table proportions for the `smallpox` data, computed by dividing each count by the table total, 6224.

GUIDED PRACTICE 3.30

G Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability.[29]

GUIDED PRACTICE 3.31

G Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of Guided Practice 3.30?[30]

GUIDED PRACTICE 3.32

G The people of Boston self-selected whether or not to be inoculated. (a) Is this study observational or was this an experiment? (b) Can we infer any causal connection using these data? (c) What are some potential confounding variables that might influence whether someone lived or died and also affect whether that person was inoculated?[31]

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

[29] $P(\texttt{result} = \texttt{died} \mid \texttt{not inoculated}) = \frac{P(\texttt{result} = \texttt{died and not inoculated})}{P(\texttt{not inoculated})} = \frac{0.1356}{0.9608} = 0.1411$.

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

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

3.2.5 General multiplication rule

Section 3.1.6 introduced the Multiplication Rule for independent processes. Here we provide the **General Multiplication Rule** for events that might not be independent.

GENERAL MULTIPLICATION RULE

If A and B represent two outcomes or events, then

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

For the term $P(A|B)$, it is useful to think of A as the outcome of interest and B as the condition.

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

EXAMPLE 3.33

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

We will compute our answer using the General Multiplication Rule and then verify it using Figure 3.12. We want to determine

$$P(\texttt{lived} \text{ and } \texttt{not inoculated})$$

(E) and we are given that

$$P(\texttt{lived} \mid \texttt{not inoculated}) = 0.8588$$
$$P(\texttt{not inoculated}) = 0.9608$$

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

$$P(\texttt{lived} \text{ and } \texttt{not inoculated}) = 0.8588 \times 0.9608 = 0.8251$$

This is equivalent to the General Multiplication Rule. We can confirm this probability in Figure 3.12 at the intersection of `no` and `lived` (with a small rounding error).

GUIDED PRACTICE 3.34

(G) Use $P(\texttt{inoculated}) = 0.0392$ and $P(\texttt{lived} \mid \texttt{inoculated}) = 0.9754$ to determine the probability that a person was both inoculated and lived.[32]

GUIDED PRACTICE 3.35

(G) If 97.54% of the inoculated people lived, what proportion of inoculated people must have died?[33]

GUIDED PRACTICE 3.36

(G) Based on the probabilities computed above, does it appear that inoculation is effective at reducing the risk of death from smallpox?[34]

[32] The answer is 0.0382, which can be verified using Figure 3.12.

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

[34] The samples are large relative to the difference in death rates for the "inoculated" and "not inoculated" groups, so it seems there is an association between `inoculated` and `outcome`. However, as noted in the solution to Guided Practice 3.32, this is an observational study and we cannot be sure if there is a causal connection. (Further research has shown that inoculation is effective at reducing death rates.)

3.2.6 Sampling without replacement

EXAMPLE 3.37

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

(E)

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

EXAMPLE 3.38

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

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

(E)

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

GUIDED PRACTICE 3.39

(G)

What rule permitted us to multiply the probabilities in Example 3.38?[35]

EXAMPLE 3.40

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

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

(E)

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

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

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

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

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

GUIDED PRACTICE 3.41

Under the setup of Example 3.40, what is the probability of being picked to answer all three questions?[36]

If we sample from a small population **without replacement**, we no longer have independence between our observations. In Example 3.38, 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 3.40, the professor sampled her students **with replacement**: she repeatedly sampled the entire class without regard to who she already picked.

GUIDED PRACTICE 3.42

Your department is holding a raffle. They sell 30 tickets and offer seven prizes. (a) They place the tickets in a hat and draw one for each prize. The tickets are sampled without replacement, i.e. the selected tickets are not placed back in the hat. What is the probability of winning a prize if you buy one ticket? (b) What if the tickets are sampled with replacement?[37]

GUIDED PRACTICE 3.43

Compare your answers in Guided Practice 3.42. How much influence does the sampling method have on your chances of winning a prize?[38]

Had we repeated Guided Practice 3.42 with 300 tickets instead of 30, we would have found something interesting: the results would be nearly identical. The probability would be 0.0233 without replacement and 0.0231 with replacement.

SAMPLING WITHOUT REPLACEMENT

When the sample size is only a small fraction of the population (under 10%), observations can be considered independent even when sampling without replacement.

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

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

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

3.2.7 Independence considerations in conditional probability

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

GUIDED PRACTICE 3.44

(G) Let X and Y represent the outcomes of rolling two dice. (a) What is the probability that the first die, X, is 1? (b) What is the probability that both X and Y are 1? (c) Use the formula for conditional probability to compute $P(Y = 1 \,|X = 1)$. (d) What is $P(Y = 1)$? Is this different from the answer from part (c)? Explain.[39]

We can show in Guided Practice 3.44(c) that the conditioning information has no influence by using the Multiplication Rule for independence processes:

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

GUIDED PRACTICE 3.45

(G) Ron is watching a roulette table in a casino and notices that the last five outcomes were `black`. He figures that the chances of getting `black` six times in a row is very small (about 1/64) and puts his paycheck on red. What is wrong with his reasoning?[40]

3.2.8 Checking for independent and mutually exclusive events

If A and B are independent events, then the probability of A being true is unchanged if B is true. Mathematically, this is written as

$$P(A|B) = P(A)$$

The General Multiplication Rule states that $P(A \text{ and } B)$ equals $P(A|B) \times P(B)$. If A and B are independent events, we can replace $P(A|B)$ with $P(A)$ and the following multiplication rule applies:

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

CHECKING WHETHER TWO EVENTS ARE INDEPENDENT

When checking whether two events A and B are independent, verify one of the following equations holds (there is no need to check both equations):

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

If the equation that is checked holds true (the left and right sides are equal), A and B are independent. If the equation does not hold, then A and B are dependent.

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

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

EXAMPLE 3.46

Are teenager college attendance and parent college degrees independent or dependent? Figure 3.13 may be helpful.

We'll use the first equation above to check for independence. If the `teen` and `parents` variables are independent, it must be true that

$$P(\texttt{teen college} \mid \texttt{parent degree}) = P(\texttt{teen college})$$

E

Using Figure 3.13, we check whether equality holds in this equation.

$$P(\texttt{teen college} \mid \texttt{parent degree}) \stackrel{?}{=} P(\texttt{teen college})$$
$$0.83 \neq 0.56$$

The value 0.83 came from a probability calculation using Figure 3.13: $\frac{231}{280} \approx 0.83$. Because the sides are not equal, teenager college attendance and parent degree are dependent. That is, we estimate the probability a teenager attended college to be higher if we know that one of the teen's parents has a college degree.

		parents		
		degree	not	Total
teen	college	231	214	445
	not	49	298	347
	Total	280	512	792

Figure 3.13: Contingency table summarizing the `family_college` data set.

GUIDED PRACTICE 3.47

G

Use the second equation in the box above to show that teenager college attendance and parent college degrees are dependent.[41]

If A and B are mutually exclusive events, then A and B cannot occur at the same time. Mathematically, this is written as

$$P(A \text{ and } B) = 0$$

The General Addition Rule states that $P(A \text{ or } B)$ equals $P(A) + P(B) - P(A \text{ and } B)$. If A and B are mutually exclusive events, we can replace $P(A \text{ and } B)$ with 0 and the following addition rule applies:

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

[41] We check for equality in the following equation:

$$P(\texttt{teen college}, \texttt{parent degree}) \stackrel{?}{=} P(\texttt{teen college}) \times P(\texttt{parent degree})$$
$$\frac{231}{792} = 0.292 \neq \frac{445}{792} \times \frac{280}{792} = 0.199$$

These terms are not equal, which confirms what we learned in Example 3.46: teenager college attendance and parent college degrees are dependent.

CHECKING WHETHER TWO EVENTS ARE MUTUALLY EXCLUSIVE (DISJOINT)

If A and B are mutually exclusive events, then they cannot occur at the same time. If asked to determine if events A and B are mutually exclusive, verify one of the following equations holds (there is no need to check both equations):

$$P(A \text{ and } B) = 0 \qquad\qquad P(A \text{ or } B) = P(A) + P(B)$$

If the equation that is checked holds true (the left and right sides are equal), A and B are mutually exclusive. If the equation does not hold, then A and B are not mutually exclusive.

EXAMPLE 3.48

Are teen college attendance and parent college degrees mutually exclusive?

Looking in the table, we see that there are 231 instances where both the teenager attended college and parents have a degree, indicating the probability of both events occurring is greater than 0. Since we have found an example where both of these events happen together, these two events are not mutually exclusive. We could more formally show this by computing the probability both events occur at the same time:

$$P(\texttt{teen college}, \texttt{parent degree}) = \frac{231}{792} \neq 0$$

Since this probability is not zero, teenager college attendance and parent college degrees are not mutually exclusive.

MUTUALLY EXCLUSIVE AND INDEPENDENT ARE DIFFERENT

If two events are mutually exclusive, then if one is true, the other cannot be true. This implies the two events are in some way connected, meaning they must be dependent.
If two events are independent, then if one occurs, it is still possible for the other to occur, meaning the events are not mutually exclusive.

DEPENDENT EVENTS NEED NOT BE MUTUALLY EXCLUSIVE.

If two events are dependent, we cannot simply conclude they are mutually exclusive. For example, the college attendance of teenagers and a college degree by one of their parents are dependent, but those events are not mutually exclusive.

3.2.9 Tree diagrams

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

The `smallpox` data fit this description. We see the population as split by `inoculation`: `yes` and `no`. Following this split, survival rates were observed for each group. This structure is reflected in the tree diagram shown in Figure 3.14. The first branch for `inoculation` is said to be the **primary** branch while the other branches are **secondary**.

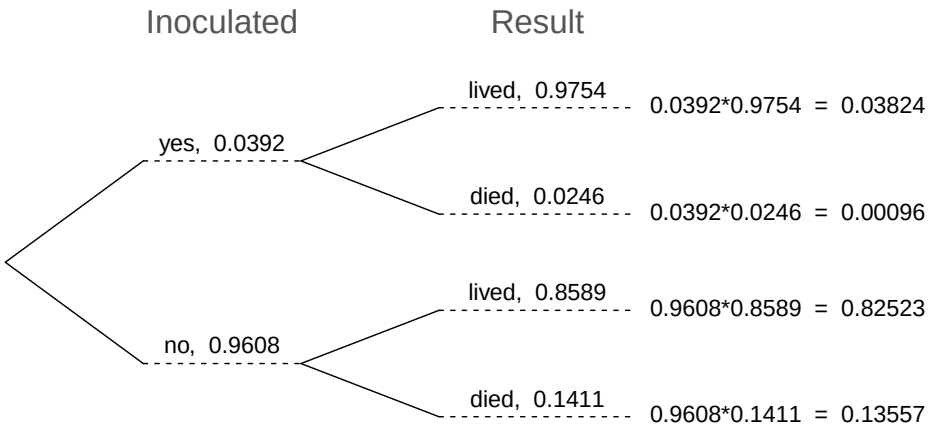


Figure 3.14: A tree diagram of the `smallpox` data set.

Tree diagrams are annotated with marginal and conditional probabilities, as shown in Figure 3.14. This tree diagram splits the smallpox data by `inoculation` into the `yes` and `no` groups with respective marginal probabilities 0.0392 and 0.9608. The secondary branches are conditioned on the first, so we assign conditional probabilities to these branches. For example, the top branch in Figure 3.14 is the probability that `lived` conditioned on the information that `inoculated`.

We may (and usually do) construct joint probabilities at the end of each branch in our tree by multiplying the numbers we come across as we move from left to right. These joint probabilities are computed using the General Multiplication Rule:

$$\begin{aligned} P(\texttt{inoculated and lived}) &= P(\texttt{inoculated}) \times P(\texttt{lived} \mid \texttt{inoculated}) \\ &= 0.0392 \times 0.9754 \\ &= 0.0382 \end{aligned}$$

EXAMPLE 3.49

What is the probability that a randomly selected person who was inoculated died?

(E)

This is equivalent to $P(\texttt{died} \mid \texttt{inoculated})$. This conditional probability can be found in the second branch as 0.0246.

EXAMPLE 3.50

What is the probability that a randomly selected person lived?

(E) There are two ways that a person could have lived: be inoculated *and* live OR not be inoculated *and* live. To find this probability, we sum the two disjoint probabilities:

$$P(\texttt{lived}) = 0.0392 \times 0.9745 + 0.9608 \times 0.8589 = 0.03824 + 0.82523 = 0.86347$$

GUIDED PRACTICE 3.51

(G) After an introductory statistics course, 78% of students can successfully construct tree diagrams. Of those who can construct tree diagrams, 97% passed, while only 57% of those students who could not construct tree diagrams passed. (a) Organize this information into a tree diagram. (b) What is the probability that a student who was able to construct tree diagrams did not pass? (c) What is the probability that a randomly selected student was able to successfully construct tree diagrams and passed? (d) What is the probability that a randomly selected student passed? [42]

3.2.10 Bayes' Theorem

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

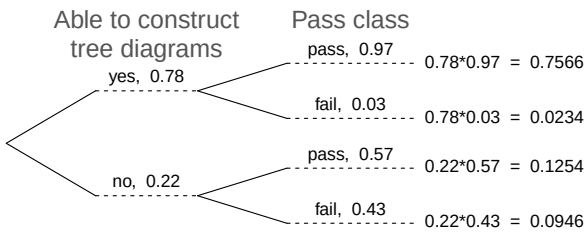
$$P(\text{statement about variable 1} \mid \text{statement about variable 2})$$

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

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

For example, instead of wanting to know $P(\text{lived} \mid \text{inoculated})$, we might want to know $P(\text{inoculated} \mid \text{lived})$. This is more challenging because it cannot be read directly from the tree diagram. In these instances we use **Bayes' Theorem**. Let's begin by looking at a new example.

[42] (a) The tree diagram is shown to the right. (b) $P(\text{not pass} \mid \text{able to construct tree diagram}) = 0.03$. (c) $P(\text{able to construct tree diagrams and passed}) = P(\text{able to construct tree diagrams}) \times P(\text{passed} \mid \text{able to construct tree diagrams}) = 0.78 \times 0.97 = 0.7566$. (d) $P(\text{passed}) = 0.7566 + 0.1254 = 0.8820$.



EXAMPLE 3.52

In Canada, about 0.35% of women over 40 will develop breast cancer in any given year. A common screening test for cancer is the mammogram, but this test is not perfect. In about 11% of patients with breast cancer, the test gives a **false negative**: it indicates a woman does not have breast cancer when she does have breast cancer. Similarly, the test gives a **false positive** in 7% of patients who do not have breast cancer: it indicates these patients have breast cancer when they actually do not.[43] If we tested a random woman over 40 for breast cancer using a mammogram and the test came back positive – that is, the test suggested the patient has cancer – what is the probability that the patient actually has breast cancer?

We are given sufficient information to quickly compute the probability of testing positive if a woman has breast cancer ($1.00 - 0.11 = 0.89$). However, we seek the inverted probability of cancer given a positive test result:

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

Here, "has BC" is an abbreviation for the patient actually having breast cancer, and "mammogram$^+$" means the mammogram screening was positive, which in this case means the test suggests the patient has breast cancer. (Watch out for the non-intuitive medical language: a *positive* test result suggests the possible presence of cancer in a mammogram screening.) We can use the conditional probability formula from the previous section: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$. Our conditional probability can be found as follows:

E

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

The probability that a mammogram is positive is as follows.

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

A tree diagram is useful for identifying each probability and is shown in Figure 3.15. Using the tree diagram, we find that

$$\begin{aligned} &P(\text{has BC} \mid \text{mammogram}^+) \\ &\quad = \frac{P(\text{has BC and mammogram}^+)}{P(\text{has BC and mammogram}^+) + P(\text{no BC and mammogram}^+)} \\ &\quad = \frac{0.0035(0.89)}{0.0035(0.89) + 0.9965(0.07)} \\ &\quad = \frac{0.00312}{0.07288} \approx 0.0428 \end{aligned}$$

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

Example 3.52 highlights why doctors often run more tests regardless of a first positive test result. When a medical condition is rare, a single positive test isn't generally definitive.

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

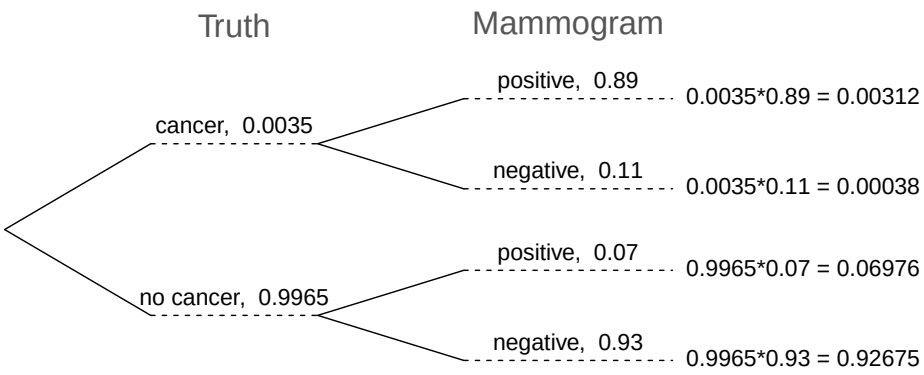


Figure 3.15: Tree diagram for Example 3.52, computing the probability a random patient who tests positive on a mammogram actually has breast cancer.

Consider again the last equation of Example 3.52. Using the tree diagram, we can see that the numerator (the top of the fraction) is equal to the following product:

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

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

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

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

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

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

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

BAYES' THEOREM: INVERTING PROBABILITIES

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

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

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

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

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

Bayes' Theorem is just a generalization of what we have done using tree diagrams. The formula need not be memorized, since it can always be derived using a tree diagram:

- The numerator identifies the probability of getting both A_1 and B.
- The denominator is the overall probability of getting B. Traverse each branch of the tree diagram that ends with event B. Add up the required products.

GUIDED PRACTICE 3.53

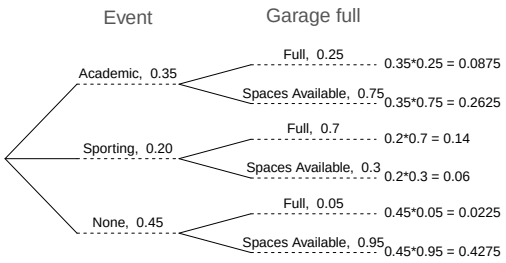
Jose visits campus every Thursday evening. However, some days the parking garage is full, often due to college events. There are academic events on 35% of evenings, sporting events on 20% of evenings, and no events on 45% of evenings. When there is an academic event, the garage fills up about 25% of 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.

G

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

$$\begin{aligned} &P(\text{sporting event}|\text{garage full}) \\ &\quad = \frac{P(\text{sporting event and garage full})}{P(\text{garage full})} \\ &\quad = \frac{0.14}{0.0875 + 0.14 + 0.0225} = 0.56. \end{aligned}$$

If the garage is full, there is a 56% probability that there is a sporting event.



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

Section summary

- A **conditional probability** can be written as $P(A|B)$ and is read, "Probability of A given B". $P(A|B)$ is the probability of A, given that B has occurred. In a conditional probability, we are given some information. In an **unconditional probability**, such as $P(A)$, we are not given any information.
- Sometimes $P(A|B)$ can be deduced. For example, when drawing without replacement from a deck of cards, $P(\text{2nd draw is an Ace} \mid \text{1st draw was an Ace}) = \frac{3}{51}$. When this is not the case, as when working with a table or a Venn diagram, one must use the conditional probability rule $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$.
- In the last section, we saw that two events are **independent** when the outcome of one has no effect on the outcome of the other. When A and B are independent, $P(A|B) = P(A)$.
- When A and B are **dependent**, find the probability of A *and* B using the **General Multiplication Rule**: $P(A \text{ and } B) = P(A|B) \times P(B)$.
- In the *special case* where A and B are **independent**, $P(A \text{ and } B) = P(A) \times P(B)$.
- If A and B are **mutually exclusive**, they must be **dependent**, since the occurrence of one of them changes the probability that the other occurs to 0.
- When sampling **without replacement**, such as drawing cards from a deck, make sure to use **conditional probabilities** when solving *and* problems.
- Sometimes, the conditional probability $P(B|A)$ may be known, but we are interested in the "inverted" probability $P(A|B)$. **Bayes' Theorem** helps us solve such conditional probabilities that cannot be easily answered. However, rather than memorize Bayes' Theorem, one can generally draw a tree diagram and apply the conditional probability rule $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$. The resulting answer often has the form $\frac{w \times x \,+\, y \times z}{w \times x}$, where w, x, y, z are numbers from a tree diagram.

Exercises

3.13 Joint and conditional probabilities. P(A) = 0.3, P(B) = 0.7

(a) Can you compute P(A and B) if you only know P(A) and P(B)?

(b) Assuming that events A and B arise from independent random processes,

i. what is P(A and B)?
ii. what is P(A or B)?
iii. what is P(A|B)?

(c) If we are given that P(A and B) = 0.1, are the random variables giving rise to events A and B independent?

(d) If we are given that P(A and B) = 0.1, what is P(A|B)?

3.14 PB & J. Suppose 80% of people like peanut butter, 89% like jelly, and 78% like both. Given that a randomly sampled person likes peanut butter, what's the probability that he also likes jelly?

3.15 Global warming. A Pew Research poll asked 1,306 Americans "From what you've read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?". The table below shows the distribution of responses by party and ideology, where the counts have been replaced with relative frequencies.[44]

		Response			
		Earth is warming	Not warming	Don't Know Refuse	Total
Party and Ideology	Conservative Republican	0.11	0.20	0.02	0.33
	Mod/Lib Republican	0.06	0.06	0.01	0.13
	Mod/Cons Democrat	0.25	0.07	0.02	0.34
	Liberal Democrat	0.18	0.01	0.01	0.20
	Total	0.60	0.34	0.06	1.00

(a) Are believing that the earth is warming and being a liberal Democrat mutually exclusive?

(b) What is the probability that a randomly chosen respondent believes the earth is warming or is a liberal Democrat?

(c) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a liberal Democrat?

(d) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a conservative Republican?

(e) Does it appear that whether or not a respondent believes the earth is warming is independent of their party and ideology? Explain your reasoning.

(f) What is the probability that a randomly chosen respondent is a moderate/liberal Republican given that he does not believe that the earth is warming?

[44] **globalWarming**.

3.16 Health coverage, relative frequencies. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table displays the distribution of health status of respondents to this survey (excellent, very good, good, fair, poor) and whether or not they have health insurance.

		Health Status					
		Excellent	Very good	Good	Fair	Poor	Total
Health Coverage	No	0.0230	0.0364	0.0427	0.0192	0.0050	0.1262
	Yes	0.2099	0.3123	0.2410	0.0817	0.0289	0.8738
	Total	0.2329	0.3486	0.2838	0.1009	0.0338	1.0000

(a) Are being in excellent health and having health coverage mutually exclusive?

(b) What is the probability that a randomly chosen individual has excellent health?

(c) What is the probability that a randomly chosen individual has excellent health given that he has health coverage?

(d) What is the probability that a randomly chosen individual has excellent health given that he doesn't have health coverage?

(e) Do having excellent health and having health coverage appear to be independent?

3.17 Burger preferences. A 2010 SurveyUSA poll asked 500 Los Angeles residents, "What is the best hamburger place in Southern California? Five Guys Burgers? In-N-Out Burger? Fat Burger? Tommy's Hamburgers? Umami Burger? Or somewhere else?" The distribution of responses by gender is shown below.[45]

		Gender		
		Male	Female	Total
Best hamburger place	Five Guys Burgers	5	6	11
	In-N-Out Burger	162	181	343
	Fat Burger	10	12	22
	Tommy's Hamburgers	27	27	54
	Umami Burger	5	1	6
	Other	26	20	46
	Not Sure	13	5	18
	Total	248	252	500

(a) Are being female and liking Five Guys Burgers mutually exclusive?

(b) What is the probability that a randomly chosen male likes In-N-Out the best?

(c) What is the probability that a randomly chosen female likes In-N-Out the best?

(d) What is the probability that a man and a woman who are dating both like In-N-Out the best? Note any assumption you make and evaluate whether you think that assumption is reasonable.

(e) What is the probability that a randomly chosen person likes Umami best or that person is female?

[45] **burgers**.

3.18 Assortative mating. Assortative mating is a nonrandom mating pattern where individuals with similar genotypes and/or phenotypes mate with one another more frequently than what would be expected under a random mating pattern. Researchers studying this topic collected data on eye colors of 204 Scandinavian men and their female partners. The table below summarizes the results. For simplicity, we only include heterosexual relationships in this exercise.[46]

		Partner (female)			
		Blue	Brown	Green	Total
Self (male)	Blue	78	23	13	114
	Brown	19	23	12	54
	Green	11	9	16	36
	Total	108	55	41	204

(a) What is the probability that a randomly chosen male respondent or his partner has blue eyes?

(b) What is the probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes?

(c) What is the probability that a randomly chosen male respondent with brown eyes has a partner with blue eyes? What about the probability of a randomly chosen male respondent with green eyes having a partner with blue eyes?

(d) Does it appear that the eye colors of male respondents and their partners are independent? Explain your reasoning.

3.19 Marbles in an urn. Imagine you have an urn containing 5 red, 3 blue, and 2 orange marbles in it.

(a) What is the probability that the first marble you draw is blue?

(b) Suppose you drew a blue marble in the first draw. If drawing with replacement, what is the probability of drawing a blue marble in the second draw?

(c) Suppose you instead drew an orange marble in the first draw. If drawing with replacement, what is the probability of drawing a blue marble in the second draw?

(d) If drawing with replacement, what is the probability of drawing two blue marbles in a row?

(e) When drawing with replacement, are the draws independent? Explain.

3.20 Socks in a drawer. In your sock drawer you have 4 blue, 5 gray, and 3 black socks. Half asleep one morning you grab 2 socks at random and put them on. Find the probability you end up wearing

(a) 2 blue socks

(b) no gray socks

(c) at least 1 black sock

(d) a green sock

(e) matching socks

3.21 Chips in a bag. Imagine you have a bag containing 5 red, 3 blue, and 2 orange chips.

(a) Suppose you draw a chip and it is blue. If drawing without replacement, what is the probability the next is also blue?

(b) Suppose you draw a chip and it is orange, and then you draw a second chip without replacement. What is the probability this second chip is blue?

(c) If drawing without replacement, what is the probability of drawing two blue chips in a row?

(d) When drawing without replacement, are the draws independent? Explain.

[46] **Laeng:2007**.

3.22 Books on a bookshelf. The table below shows the distribution of books on a bookcase based on whether they are nonfiction or fiction and hardcover or paperback.

		Format		
		Hardcover	Paperback	Total
Type	Fiction	13	59	72
	Nonfiction	15	8	23
	Total	28	67	95

(a) Find the probability of drawing a hardcover book first then a paperback fiction book second when drawing without replacement.

(b) Determine the probability of drawing a fiction book first and then a hardcover book second, when drawing without replacement.

(c) Calculate the probability of the scenario in part (b), except this time complete the calculations under the scenario where the first book is placed back on the bookcase before randomly drawing the second book.

(d) The final answers to parts (b) and (c) are very similar. Explain why this is the case.

3.23 Student outfits. In a classroom with 24 students, 7 students are wearing jeans, 4 are wearing shorts, 8 are wearing skirts, and the rest are wearing leggings. If we randomly select 3 students without replacement, what is the probability that one of the selected students is wearing leggings and the other two are wearing jeans? Note that these are mutually exclusive clothing options.

3.24 The birthday problem. Suppose we pick three people at random. For each of the following questions, ignore the special case where someone might be born on February 29th, and assume that births are evenly distributed throughout the year.

(a) What is the probability that the first two people share a birthday?

(b) What is the probability that at least two people share a birthday?

3.25 Drawing box plots. After an introductory statistics course, 80% of students can successfully construct box plots. Of those who can construct box plots, 86% passed, while only 65% of those students who could not construct box plots passed.

(a) Construct a tree diagram of this scenario.

(b) Calculate the probability that a student is able to construct a box plot if it is known that he passed.

3.26 Predisposition for thrombosis. A genetic test is used to determine if people have a predisposition for *thrombosis*, which is the formation of a blood clot inside a blood vessel that obstructs the flow of blood through the circulatory system. It is believed that 3% of people actually have this predisposition. The genetic test is 99% accurate if a person actually has the predisposition, meaning that the probability of a positive test result when a person actually has the predisposition is 0.99. The test is 98% accurate if a person does not have the predisposition. What is the probability that a randomly selected person who tests positive for the predisposition by the test actually has the predisposition?

3.27 It's never lupus. Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed tha t 2% of the population suffer from this disease. The test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease. There is a line from the Fox television show *House* that is often used after a patient tests positive for lupus: "It's never lupus." Do you think there is truth to this statement? Use appropriate probabilities to support your answer.

3.28 Exit poll. Edison Research gathered exit poll results from several sources for the Wisconsin recall election of Scott Walker. They found that 53% of the respondents voted in favor of Scott Walker. Additionally, they estimated that of those who did vote in favor for Scott Walker, 37% had a college degree, while 44% of those who voted against Scott Walker had a college degree. Suppose we randomly sampled a person who participated in the exit poll and found that he had a college degree. What is the probability that he voted in favor of Scott Walker?[47]

[47] **data:scott**.

3.3 Simulations

What is the probability of getting a sum greater than 16 in three rolls of a die? Finding all possible combinations that satisfy this would be tedious, but we could conduct a physical simulation or a computer simulation to estimate this probability. With modern computing power, simulations have become an important and powerful tool for data scientists. In this section, we will look at the concepts that underlie simulations.

Learning objectives

1. Understand the purpose of a simulation and recognize the application of the long-run relative frequency interpretation of probability.
2. Understand how random digit tables work and how to assign digits to outcomes.
3. Be able to repeat a simulation a set number of trials or until a condition is true, and use the results to estimate the probability of interest.

3.3.1 Setting up and carrying out simulations

In the previous section we saw how to apply the binomial formula to find the probability of exactly x successes in n independent trials when a success has probability p. Sometimes we have a problem we want to solve but we don't know the appropriate formula, or even worse, a formula may not exist. In this case, one common approach is to estimate the probability using **simulations**.

You may already be familiar with simulations. Want to know the probability of rolling a sum of 7 with a pair of dice? Roll a pair of dice many, many, many times and see what proportion of times the sum is 7. The more times you roll the pair of dice, the better the estimate will tend to be. Of course, such experiments can be time consuming or even infeasible.

In this section, we consider simulations using **random numbers**. Random numbers (or technically, *psuedo-random numbers*) can be produced using a calculator or computer. Random digits are produced such that each digit, *0-9*, is equally likely to come up in each spot. You'll find that occasionally we may have the same number in a row – sometimes multiple times – but in the long run, each digit should appear 1/10th of the time.

	Column			
Row	1-5	6-10	11-15	16-20
1	43087	41864	51009	39689
2	63432	72132	40269	56103
3	19025	83056	62511	52598
4	85117	16706	31083	24816
5	16285	56280	01494	90240
6	94342	18473	50845	77757
7	61099	14136	39052	50235
8	37537	58839	56876	02960
9	04510	16172	90838	15210
10	27217	12151	52645	96218

Figure 3.16: Random number table. A full page of random numbers may be found in Appendix C.1 on page 263.

EXAMPLE 3.54

Mika's favorite brand of cereal is running a special where 20% of the cereal boxes contain a prize. Mika really wants that prize. If her mother buys 6 boxes of the cereal over the next few months, what is the probability Mika will get a prize?

To solve this problem using simulation, we need to be able to assign digits to outcomes. Each box should have a 20% chance of having a prize and an 80% chance of not having a prize. Therefore, a valid assignment would be:

0, 1 → prize

2-9 → no prize

(E) Of the ten possible digits (*0, 1, 2, ..., 8, 9*), two of them, i.e. 20% of them, correspond to winning a prize, which exactly matches the odds that a cereal box contains a prize.

In Mika's simulation, one trial will consist of 6 boxes of cereal, and therefore a trial will require six digits (each digit will correspond to one box of cereal). We will repeat the simulation for 20 trials. Therefore we will need 20 sets of 6 digits. Let's begin on row 1 of the random digit table, shown in Figure 3.16. If a trial consisted of 5 digits, we could use the first 5 digits going across: *43087*. Because here a trial consists of 6 digits, it may be easier to read down the table, rather than read across. We will let trial 1 consist of the first 6 digits in column 1 (*461819*), trial 2 consist of the first 6 digits in column 2 (*339564*), etc. For this simulation, we will end up using the first 6 rows of each of the 20 columns.

In trial 1, there are two *1*'s, so we record that as a success; in this trial there were actually two prizes. In trial 2 there were no *0*'s or *1*'s, therefore we do not record this as a success. In trial 3 there were three prizes, so we record this as a success. The rest of this exercise is left as a Guided Practice problem for you to complete.

GUIDED PRACTICE 3.55

G Finish the simulation above and report the estimate for the probability that Mika will get a prize if her mother buys 6 boxes of cereal where each one has a 20% chance of containing a prize.[48]

GUIDED PRACTICE 3.56

G In the previous example, the probability that a box of cereal contains a prize is 20%. The question presented is equivalent to asking, what is the probability of getting at least one prize in six randomly selected boxes of cereal. This probability question can be solved explicitly using the method of complements. Find this probability. How does the estimate arrived at by simulation compare to this probability?[49]

We can also use simulations to estimate quantities other than probabilities. Consider the following example.

EXAMPLE 3.57

Let's say that instead of buying exactly 6 boxes of cereal, Mika's mother agrees to buy boxes of this cereal *until* she finds one with a prize. On average, how many boxes of cereal would one have to buy until one gets a prize?

E For this question, we can use the same digit assignment. However, our stopping rule is different. Each trial may require a different number of digits. For each trial, the stopping rule is: look at digits until we encounter a *0* or a *1*. Then, record how many digits/boxes of cereal it took. Repeat the simulation for 20 trials, and then average the numbers from each trial.

Let's begin again at row 1. We can read across or down, depending upon what is most convenient. Since there are 20 columns and we want 20 trials, we will read down the columns. Starting at column 1, we count how many digits (boxes of cereal) we encounter until we reach a *0* or *1* (which represent a prize). For trial 1 we see *461*, so we record 3. For trial 2 we see *3395641*, so we record 7. For trial 3, we see *0*, so we record 1. The rest of this exercise is left as a Guided Practice problem for you to complete.

GUIDED PRACTICE 3.58

G Finish the simulation above and report your estimate for the average number of boxes of cereal one would have to buy until encountering a prize, where the probability of a prize in each box is 20%.[50]

[48] The trials that contain at least one 0 or 1 and therefore are successes are trials: 1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, and 20. There were 17 successes among the 20 trials, so our estimate of the probability based on this simulation is $17/20 = 0.85$.

[49] The true probability is given by $1 - P(\text{no prizes in six boxes}) = 1 - 0.8^6 = 0.74$. The estimate arrived at by simulation was 11% too high. Note: We only repeated the simulation 20 times. If we had repeated it 1000 times, we would (very likely) have gotten an estimate closer to the true probability.

[50] For the 20 trials, the number of digits we see until we encounter a *0* or *1* is: 3,7,1,4,9, 4,1,2,4,5, 5,1,1,1,3, 8,5,2,2,6. Now we take the average of these 20 numbers to get $74/20 = 3.7$.

EXAMPLE 3.59

Now, consider a case where the probability of interest is not 20%, but rather 28%. Which digits should correspond to success and which to failure?

This example is more complicated because with only 10 digits, there is no way to select exactly 28% of them. Therefore, each observation will have to consist of *two* digits. We can use two digits at a time and assign pairs of digits as follows:

00-27 → success
28-99 → failure

GUIDED PRACTICE 3.60

Assume the probability of winning a particular casino game is 45%. We want to carry out a simulation to estimate the probability that we will win at least 5 times in 10 plays. We will use 30 trials of the simulation. Assign digits to outcomes. Also, how many total digits will we require to run this simulation?[51]

GUIDED PRACTICE 3.61

Assume carnival spinner has 7 slots. We want to carry out a simulation to estimate the probability that we will win at least 10 times in 60 plays. Repeat 100 trials of the simulation. Assign digits to outcomes. Also, how many total digits will we require to run this simulation?[52]

Does anyone perform simulations like this? Sort of. Simulations are used a lot in statistics, and these often require the same principles covered in this section to properly set up those simulations. The difference is in implementation after the setup. Rather than use a random number table, a data scientist will write a program that uses a pseudo-random number generator in a computer to run the simulations very quickly – often times millions of trials each second, which provides much more accurate estimates than running a couple dozen trials by hand.

[51] One possible assignment is: *00-44* → win and *45-99* → lose. Each trial requires 10 pairs of digits, so we will need 30 sets of 10 pairs of digits for a total of $30 \times 10 \times 2 = 600$ digits.

[52] Note that $1/7 = 0.142857...$ This makes it tricky to assign digits to outcomes. The best approach here would be to exclude some of the digits from the simulation. We can assign *0* to success and *1-6* to failure. This corresponds to a 1/7 chance of getting a success. If we encounter a *7*, *8*, or *9*, we will just skip over it. Because we don't know how many *7*, *8*, or *9*'s we will encounter, we do not know how many total digits we will end up using for the simulation. (If you want a challenge, try to estimate the total number of digits you would need.)

Section summary

- When a probability is difficult to determine via a formula, one can set up a **simulation** to estimate the probability.
- The **relative frequency** theory of probability and the **Law of Large Numbers** are the mathematical underpinning of simulations. A larger number of trials should tend to produce better estimates.
- The first step to setting up a simulation is to assign digits to represent outcomes. This should be done in such a way as to give the event of interest the correct probability. Then, using a random number table, calculator, or computer, generate random digits (outcomes). Repeat this a specified number of trials or until a given stopping rule. When this is finished, count up how many times the event happened and divide that by the number of trials to get the estimate of the probability.

Exercises

3.29 Smog check, Part I . Suppose 16% of cars fail pollution tests (smog checks) in California. We would like to estimate the probability that an entire fleet of seven cars would pass using a simulation. We assume each car is independent. We only want to know if the entire fleet passed, i.e. none of the cars failed. What is wrong with each of the following simulations to represent whether an entire (simulated) fleet passed?

(a) Flip a coin seven times where each toss represents a car. A head means the car passed and a tail means it failed. If all cars passed, we report PASS for the fleet. If at least one car failed, we report FAIL.

(b) Read across a random number table starting at line 5. If a number is a 0 or 1, let it represent a failed car. Otherwise the car passes. We report PASS if all cars passed and FAIL otherwise.

(c) Read across a random number table, looking at two digits for each simulated car. If a pair is in the range [00-16], then the corresponding car failed. If it is in [17-99], the car passed. We report PASS if all cars passed and FAIL otherwise.

3.30 Left-handed. Studies suggest that approximately 10% of the world population is left-handed. Use ten simulations to answer each of the following questions. For each question, describe your simulation scheme clearly.

(a) What is the probability that at least one out of eight people are left-handed?

(b) On average, how many people would you have to sample until the first person who is left-handed?

(c) On average, how many left-handed people would you expect to find among a random sample of six people?

3.31 Smog check, Part II. Consider the fleet of seven cars in Exercise **??**. Remember that 16% of cars fail pollution tests (smog checks) in California, and that we assume each car is independent.

(a) Write out how to calculate the probability of the fleet failing, i.e. at least one of the cars in the fleet failing, via simulation.

(b) Simulate 5 fleets. Based on these simulations, estimate the probability at least one car will fail in a fleet.

(c) Compute the probability at least one car fails in a fleet of seven.

3.32 To catch a thief. Suppose that at a retail store, $1/5^{th}$ of all employees steal some amount of merchandise. The stores would like to put an end to this practice, and one idea is to use lie detector tests to catch and fire thieves. However, there is a problem: lie detectors are not 100% accurate. Suppose it is known that a lie detector has a failure rate of 25%. A thief will slip by the test 25% of the time and an honest employee will only pass 75% of the time.

(a) Describe how you would simulate whether an employee is honest or is a thief using a random number table. Write your simulation very carefully so someone else can read it and follow the directions exactly.

(b) Using a random number table, simulate 20 employees working at this store and determine if they are honest or not. Make sure to record the random digits assigned to each employee as you will refer back to these in part (c).

(c) Determine the result of the lie detector test for each simulated employee from part (b) using a new simulation scheme.

(d) How many of these employees are "honest and passed" and how many are "honest and failed"?

(e) How many of these employees are "thief and passed" and how many are "thief and failed"?

(f) Suppose the management decided to fire everyone who failed the lie detector test. What percent of fired employees were honest? What percent of not fired employees were thieves?

3.4 Random variables

The chance of landing on single number in the game of roulette is 1/38 and the pay is 35:1. The chance of landing on Red is 18/38 and the pay is 1:1. Which game has the higher expected value? The higher standard deviation of expected winnings? How do we interpret these quantities in this context? If you were to play each game 20 times, what would the *distribution* of possible outcomes look like? In this section, we define and summarize random variables such as this, and we look at some of their properties.

Learning objectives

1. Define a probability distribution and what makes a distribution a valid probability distribution.
2. Summarize a discrete probability distribution graphically using a histogram and verbally with respect to center, spread, and shape.
3. Calculate and interpret the mean (expected value) and standard deviation of a random variable.
4. Calculate the mean and standard deviation of a transformed random variable.
5. Calculate the mean of the sum or difference of random variables.
6. Calculate the standard deviation of the sum or difference of random variables when those variables are independent.

3.4.1 Introduction to expected value

EXAMPLE 3.62

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

(E)

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

GUIDED PRACTICE 3.63

(G) Would you be surprised if the bookstore sold slightly more or less than 105 books?[53]

[53] 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 3.64

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

(E)

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

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

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

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

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

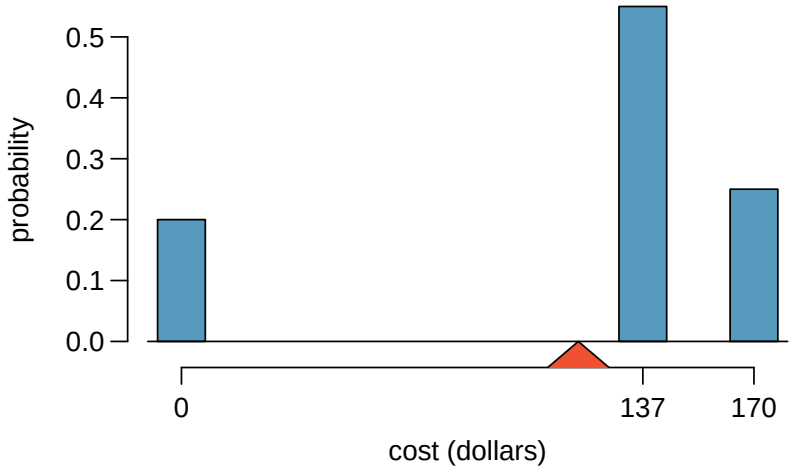


Figure 3.17: Probability distribution for the bookstore's revenue from one student. The triangle represents the average revenue per student.

EXAMPLE 3.65

What is the average revenue per student for this course?

(E)

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

3.4.2 Probability distributions

A **probability distribution** is a table of all disjoint outcomes and their associated probabilities. Figure 3.18 shows the probability distribution for the sum of two dice.

RULES FOR PROBABILITY DISTRIBUTIONS

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

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

GUIDED PRACTICE 3.66

Figure 3.19 suggests three distributions for household income in the United States. Only one is correct. Which one must it be? What is wrong with the other two?[54]

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

Figure 3.18: Probability distribution for the sum of two dice.

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

Figure 3.19: Proposed distributions of US household incomes (Guided Practice 3.66).

Chapter 2 emphasized the importance of plotting data to provide quick summaries. Probability distributions can also be summarized in a histogram or bar plot. The probability distribution for the sum of two dice is shown in Figure 3.18 and its histogram is plotted in Figure 3.20. The distribution of US household incomes is shown in Figure 3.21 as a bar plot. The presence of the 100+ category makes it difficult to represent it with a regular histogram.[55]

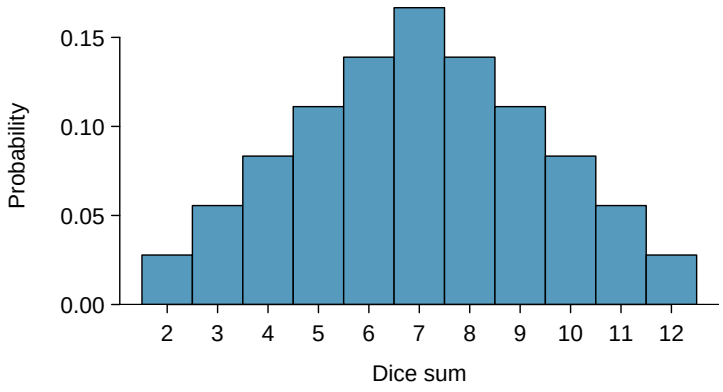


Figure 3.20: A histogram for the probability distribution of the sum of two dice.

In these bar plots, the bar heights represent the probabilities of outcomes. If the outcomes are numerical and discrete, it is usually (visually) convenient to make a histogram, as in the case of the sum of two dice. Another example of plotting the bars at their respective locations is shown in Figure 3.17.

[54]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).

[55]It is also possible to construct a distribution plot when income is not artificially binned into four groups. Density histograms for *continuous* distributions are considered in Section **??**.

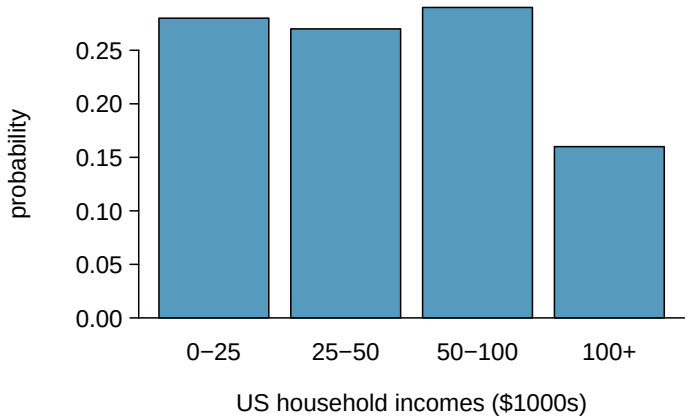


Figure 3.21: A bar graph for the probability distribution of US household income. Because it is artificially separated into four unequal bins, this graph fails to show the shape or skew of the distribution.

3.4.3 Expectation

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

RANDOM VARIABLE

A random process or variable with a numerical outcome.

The possible outcomes of X are labeled with a corresponding lower case letter x and subscripts. For example, we write $x_1 = \$0$, $x_2 = \$137$, and $x_3 = \$170$, which occur with probabilities 0.20, 0.55, and 0.25. The distribution of X is summarized in Figure 3.17 and Figure 3.22.

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

Figure 3.22: The probability distribution for the random variable X, representing the bookstore's revenue from a single student. We use $P(x_i)$ to represent the probability of x_i.

We computed the average outcome of X as \$117.85 in Example 3.65. We call this average the **expected value** of X, denoted by $E(X)$. The expected value of a random variable is computed by adding each outcome weighted by its probability:

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

EXPECTED VALUE OF A DISCRETE RANDOM VARIABLE

If X takes outcomes x_1, x_2, ..., x_n with probabilities $P(x_1)$, $P(x_2)$, ..., $P(x_n)$, the mean, or expected value, of X is the sum of each outcome multiplied by its corresponding probability:

$$\begin{aligned} \mu_X = E(X) &= x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + \cdots + x_n \cdot P(x_n) \\ &= \sum_{i=1}^{n} x_i \cdot P(x_i) \end{aligned}$$

The expected value for a random variable represents the average outcome. For example, $E(X) = 117.85$ represents the average amount the bookstore expects to make from a single student, which we could also write as $\mu = 117.85$. While the bookstore will make more than this on some students and less than this on other students, the average of many randomly selected students will be near \$117.85.

It is also possible to compute the expected value of a continuous random variable (see Section **??**). However, it requires a little calculus and we save it for a later class.[56]

In physics, the expectation holds the same meaning as the center of gravity. The distribution can be represented by a series of weights at each outcome, and the mean represents the balancing point. This is represented in Figures 3.17 and 3.23. The idea of a center of gravity also expands to continuous probability distributions. Figure 3.24 shows a continuous probability distribution balanced atop a wedge placed at the mean.

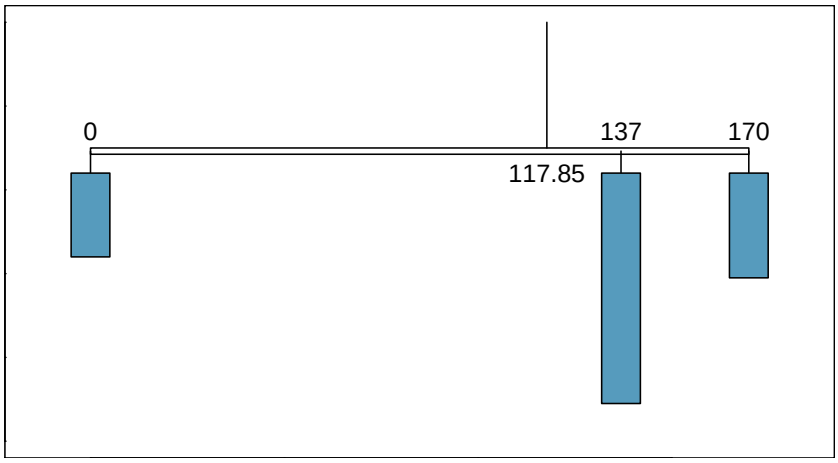


Figure 3.23: A weight system representing the probability distribution for X. The string holds the distribution at the mean to keep the system balanced.

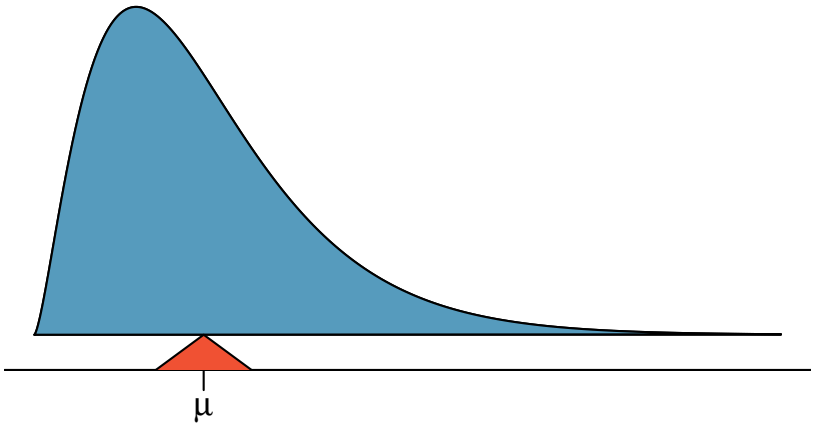


Figure 3.24: A continuous distribution can also be balanced at its mean.

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

3.4.4 Variability in random variables

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

The variance and standard deviation can be used to describe the variability of a random variable. Section 2.2.2 introduced a method for finding the variance and standard deviation for a data set. We first computed deviations from the mean $(x_i - \mu)$, squared those deviations, and took an average to get the variance. In the case of a random variable, we again compute squared deviations. However, we take their sum weighted by their corresponding probabilities, just like we did for the expectation. This weighted sum of squared deviations equals the variance, and we calculate the standard deviation by taking the square root of the variance, just as we did in Section 2.2.2.

VARIANCE AND STANDARD DEVIATION OF A DISCRETE RANDOM VARIABLE

If X takes outcomes x_1, x_2, ..., x_n with probabilities $P(x_1)$, $P(x_2)$, ..., $P(x_n)$ and expected value $\mu_X = E(X)$, then to find the standard deviation of X, we first find the variance and then take its square root.

$$\begin{aligned} Var(X) = \sigma_x^2 &= (x_1 - \mu_x)^2 \cdot P(x_1) + (x_2 - \mu_x)^2 \cdot P(x_2) + \cdots + (x_n - \mu_x)^2 \cdot P(x_n) \\ &= \sum_{i=1}^{n} (x_i - \mu_x)^2 \cdot P(x_i) \\ SD(X) = \sigma_x &= \sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2 \cdot P(x_i)} \end{aligned}$$

Just as it is possible to compute the mean of a continuous random variable using calculus, we can also use calculus to compute the variance.[57] However, this topic is beyond the scope of the AP exam.

EXAMPLE 3.67

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

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

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

(E)

Thus, the expected value is $\mu_X = 117.85$, which we computed earlier. The variance can be constructed using a similar table:

i	1	2	3	Total
x_i	\$0	\$137	\$170	
$P(x_i)$	0.20	0.55	0.25	
$x_i - \mu_X$	-117.85	19.15	52.15	
$(x_i - \mu_X)^2$	13888.62	366.72	2719.62	
$(x_i - \mu_X)^2 \cdot P(x_i)$	2777.7	201.7	679.9	3659.3

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

[57] $\sigma_x^2 = \int (x - \mu_X)^2 f(x) dx$ where $f(x)$ represents a function for the density curve.

GUIDED PRACTICE 3.68

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.[58]

(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.

3.4.5 Linear transformations of a random variable

An online store is selling a limited edition t-shirt. The maximum a person is allowed to buy is 3. Let X be a random variable that represents how many of the t-shirts a t-shirt buyer orders. The probability distribution of X is given in the following table.

x_i	1	2	3
$P(x_i)$	0.6	0.3	0.1

Using the methods of the previous section we can find that the mean $\mu_X = 1.5$ and the standard deviation $\sigma_X = 0.67$. Suppose that the cost of each t-shirt is $30 and that there is flat rate $5 shipping fee. The amount of money a t-shirt buyer pays, then, is $30X + 5$, where X is the number of t-shirts ordered. To calculate the mean and standard deviation for the amount of money a t-shirt buyers pays, we could define a new variable Y as follows:

$$Y = 30X + 5$$

GUIDED PRACTICE 3.69

Verify that the distribution of Y is given by the table below.[59]

y_i	$35	$65	$95
$P(y_i)$	0.6	0.3	0.1

[58] (a) 100% - 25% - 60% = 15% of students do not buy any books for the class. Part (b) is represented by the first two lines in the table below. The expectation for part (c) is given as the total on the line $y_i \cdot P(y_i)$. The result of part (d) is the square-root of the variance listed on in the total on the last line: $\sigma_Y = \sqrt{Var(Y)} = \sqrt{4800} = 69.28$.

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

[59] $30 \times 1 + 5 = 35$; $30 \times 2 + 5 = 65$; $30 \times 3 + 5 = 95$

Using this new table, we can compute the mean and standard deviation of the cost for t-shirt orders. However, because Y is a linear transformation of X, we can use the properties from Section 2.2.7. Recall that multiplying every X by 30 multiplies both the mean and standard deviation by 30. Adding 5 only adds 5 to the mean, not the standard deviation. Therefore,

$$\begin{aligned}\mu_{30X+5} &= E(30X+5) \\ &= 30 \times E(X) + 5 \\ &= 30 \times 1.5 + 5 \\ &= 45.00\end{aligned} \qquad \begin{aligned}\sigma_{30X+5} &= SD(30X+5) \\ &= 30 \times SD(X) \\ &= 30 \times 0.67 \\ &= 20.10\end{aligned}$$

Among t-shirt buyers, they spend an average of \$45.00, with a standard deviation of \$20.10.

LINEAR TRANSFORMATIONS OF A RANDOM VARIABLE

If X is a random variable, then a linear transformation is given by $aX + b$, where a and b are some fixed numbers.

$$E(aX+b) = a \times E(X) + b \qquad SD(aX+b) = |a| \times SD(X)$$

3.4.6 Linear combinations of random variables

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

EXAMPLE 3.70

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

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

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

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

EXAMPLE 3.71

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

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

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

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

GUIDED PRACTICE 3.72

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.[60]

GUIDED PRACTICE 3.73

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?[61]

GUIDED PRACTICE 3.74

Would you be surprised if John's weekly commute wasn't exactly 90 minutes or if Elena didn't make exactly \$152? Explain.[62]

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

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

$$aX + bY$$

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

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

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

When considering the average of a linear combination of random variables, it is safe to plug in the mean of each random variable and then compute the final result. For a few examples of nonlinear combinations of random variables – cases where we cannot simply plug in the means – see the footnote.[63]

LINEAR COMBINATIONS OF RANDOM VARIABLES AND THE AVERAGE RESULT

If X and Y are random variables, then a linear combination of the random variables is given by $aX + bY$, where a and b are some fixed numbers. To compute the average value of a linear combination of random variables, plug in the average of each individual random variable and compute the result:

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

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

[60] She will make X dollars on the TV but spend Y dollars on the toaster oven: $X - Y$.

[61] $E(X - Y) = E(X) - E(Y) = 175 - 23 = \$152$. She should expect to make about \$152.

[62] 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.

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

EXAMPLE 3.75

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.

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

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

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

GUIDED PRACTICE 3.76

(G) 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.[64]

GUIDED PRACTICE 3.77

(G) You should have found that Leonard expects a positive gain in Guided Practice 3.76. However, would you be surprised if he actually had a loss this month?[65]

3.4.7 Variability in linear combinations of random variables

Quantifying the average outcome from a linear combination of random variables is helpful, but it is also important to have some sense of the uncertainty associated with the total outcome of that combination of random variables. The expected net gain or loss of Leonard's stock portfolio was considered in Guided Practice 3.76. However, there was no quantitative discussion of the volatility of this portfolio. For instance, while the average monthly gain might be about \$134 according to the data, that gain is not guaranteed. Figure 3.25 shows the monthly changes in a portfolio like Leonard's during the 36 months from 2009 to 2011. The gains and losses vary widely, and quantifying these fluctuations is important when investing in stocks.

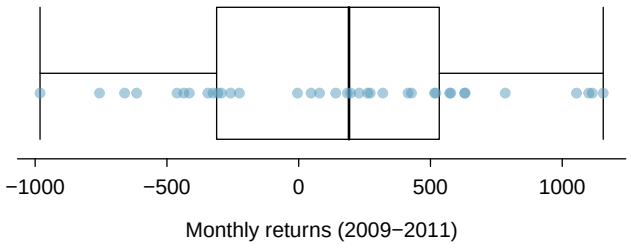


Figure 3.25: The change in a portfolio like Leonard's for the 36 months from 2009 to 2011, where \$6000 is in Google's stock and \$2000 is in Exxon Mobil's.

Just as we have done in many previous cases, we use the variance and standard deviation to describe the uncertainty associated with Leonard's monthly returns. To do so, the standard deviations and variances of each stock's monthly return will be useful, and these are shown in Figure 3.26. The stocks' returns are nearly independent.

[64] $E(\$6000 \times X + \$2000 \times Y) = \$6000 \times 0.021 + \$2000 \times 0.004 = \$134$.

[65] No. While stocks tend to rise over time, they are often volatile in the short term.

	Mean ($\bar{x}$)	Standard deviation (s)	Variance (s^2)
GOOG	0.0210	0.0849	0.0072
XOM	0.0038	0.0520	0.0027

Figure 3.26: The mean, standard deviation, and variance of the GOOG and XOM stocks. These statistics were estimated from historical stock data, so notation used for sample statistics has been used.

We want to describe the uncertainty of Leonard's monthly returns by finding the standard deviation of the return on his combined portfolio. First, we note that the variance of a sum has a nice property: the variance of a sum is the sum of the variances. That is, if X and Y are independent random variables:

$$Var(X+Y) = Var(X) + Var(Y)$$

Because the standard deviation is the square root of the variance, we can rewrite this equation using standard deviations:

$$(SD_{X+Y})^2 = (SD_X)^2 + (SD_Y)^2$$

This equation might remind you of a theorem from geometry: $c^2 = a^2 + b^2$. The equation for the standard deviation of the sum of two independent random variables looks analogous to the Pythagorean Theorem. Just as the Pythagorean Theorem only holds for right triangles, this equation only holds when X and Y are *independent*.[66]

STANDARD DEVIATION OF THE SUM AND DIFFERENCE OF RANDOM VARIABLES

If X and Y are *independent* random variables:

$$SD_{X+Y} = SD_{X-Y} = \sqrt{(SD_X)^2 + (SD_Y)^2}$$

Because $SD_Y = SD_{-Y}$, the standard deviation of the difference of two variables equals the standard deviation of the sum of two variables. This property holds for more than two variables as well. For example, if X, Y, and Z are independent random variables:

$$SD_{X+Y+Z} = SD_{X-Y-Z} = \sqrt{(SD_X)^2 + (SD_Y)^2 + (SD_Z)^2}$$

If we need the standard deviation of a linear combination of independent variables, such as $aX + bY$, we can consider aX and bY as two new variables. Recall that multiplying all of the values of variable by a positive constant multiplies the standard deviation by that constant. Thus, $SD_{aX} = a \times SD_X$ and $SD_{bY} = b \times SD_Y$. It follows that:

$$SD_{aX+bY} = \sqrt{(a \times SD_X)^2 + (b \times SD_Y)^2}$$

This equation can be used to compute the standard deviation of Leonard's monthly return. Recall that Leonard has \$6,000 in Google stock and \$2,000 in Exxon Mobil's stock. From Figure 3.26, the standard deviation of Google stock is 0.0849 and the standard deviation of Exxon Mobile stock is 0.0520.

$$\begin{aligned} SD_{6000X+2000Y} &= \sqrt{(6000 \times SD_X)^2 + (2000 \times SD_Y)^2} \\ &= \sqrt{(6000 \times 0.0849)^2 + (2000 \times .0520)^2} \\ &= \sqrt{270{,}304} = 520 \end{aligned}$$

The standard deviation of the total is \$520. While an average monthly return of \$134 on an \$8000 investment is nothing to scoff at, the monthly returns are so volatile that Leonard should not expect this income to be very stable.

[66] Another word for independent is orthogonal, meaning right angle! When X and Y are dependent, the equation for SD_{X+Y} becomes analogous to the law of cosines.

STANDARD DEVIATION OF LINEAR COMBINATIONS OF RANDOM VARIABLES

To find the standard deviation of a linear combination of random variables, we first consider aX and bY separately. We find the standard deviation of each, and then we apply the equation for the standard deviation of the sum of two variables:

$$SD_{aX+bY} = \sqrt{(a \times SD_X)^2 + (b \times SD_Y)^2}$$

This equation is valid as long as the random variables X and Y are *independent* of each other.

EXAMPLE 3.78

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

The expression for John's commute time is

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

(E)

Each coefficient is 1, so the standard deviation of the total weekly commute time is

$$\begin{aligned} \text{SD} &= \sqrt{(1 \times 4)^2 + (1 \times 4)^2 + (1 \times 4)^2 + (1 \times 4)^2 + (1 \times 4)^2} \\ &= \sqrt{5 \times (4)^2} \\ &= 8.94 \end{aligned}$$

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

GUIDED PRACTICE 3.79

(G) The computation in Example 3.78 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.[67]

GUIDED PRACTICE 3.80

(G) Consider Elena's two auctions from Guided Practice 3.72 on page 135. 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.[68]

Consider again Guided Practice 3.80. 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.

3.4.8 Normal approximation for sums of random variables

We have seen that many distributions are approximately normal. The sum and the difference of normally distributed variables is also normal. While we cannot prove this here, the usefulness of it is seen in the following example.

[67] 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.

[68] The equation for Elena can be written as: $(1) \times X + (-1) \times Y$. To find the SD of this new variable we do:

$$SD_{(1)\times X+(-1)\times Y} = \sqrt{(1 \times SD_X)^2 + (-1 \times SD_Y)^2 = (1 \times 25)^2 + (-1 \times 8)^2} = 26.25$$

The SD is about \$26.25.

EXAMPLE 3.81

Three friends are playing a cooperative video game in which they have to complete a puzzle as fast as possible. Assume that the individual times of the 3 friends are independent of each other. The individual times of the friends in similar puzzles are approximately normally distributed with the following means and standard deviations.

	Mean	SD
Friend 1	5.6	0.11
Friend 2	5.8	0.13
Friend 3	6.1	0.12

To advance to the next level of the game, the friends' total time must not exceed 17.1 minutes. What is the probability that they will advance to the next level?

Because each friend's time is approximately normally distributed, *the sum of their times is also approximately normally distributed*. We will do a normal approximation, but first we need to find the mean and standard deviation of the *sum*. We learned how to do this in Section 3.4.

Let the three friends be labeled X, Y, Z. We want $P(X+Y+Z<17.1)$. The mean and standard deviation of the sum of X, Y, and Z is given by:

(E)

$$\begin{aligned}\mu_{sum} &= E(X+Y+Z)\\ &= E(X)+E(Y)+E(Z)\\ &= 4.6+4.8+4.5\\ &= 17.5\end{aligned}$$

$$\begin{aligned}\sigma_{sum} &= \sqrt{(SD_X)^2+(SD_Y)^2+(SD_Z)^2}\\ &= \sqrt{(0.11)^2+(0.13)^2+(0.12)^2}\\ &= 0.208\end{aligned}$$

Now we can find the Z-score.

$$\begin{aligned}Z &= \frac{x_{sum}-\mu_{sum}}{\sigma_{sum}}\\ &= \frac{17.1-17.5}{0.208}\\ &= -1.92\end{aligned}$$

Finally, we want the probability that the sum is less than 17.5, so we shade the area to the left of $Z=-1.92$. Using technology, we get

$$P(Z<-1.92)=0.027$$

There is a 2.7% chance that the friends will advance to the next level.

GUIDED PRACTICE 3.82

(G) What is the probability that Friend 2 will complete the puzzle with a faster time than Friend 1? Hint: find $P(Y<X)$, or $P(Y-X<0)$.[69]

[69] First find the mean and standard deviation of $Y-X$. The mean of $Y-X$ is $\mu_{Y-X}=5.8-5.6=0.2$. The standard deviation is $SD_{Y-X}=\sqrt{(0.13)^2+(0.11)^2}=0.170$. Then $Z=\frac{0-0.2}{0.170}=-1.18$ and $P(Z<-1.18)=.119$. There is an 11.9% chance that Friend 2 will complete the puzzle with a faster time than Friend 1.

Section summary

- A **discrete probability distribution** can be summarized in a table that consists of all possible outcomes of a random variable and the probabilities of those outcomes. The outcomes must be disjoint, and the sum of the probabilities must equal 1.
- A probability distribution can be represented with a histogram and, like the distributions of data that we saw in Chapter 2, can be summarized by its **center**, **spread**, and **shape**.
- When given a probability distribution table, we can calculate the **mean** (expected value) and **standard deviation** of a random variable using the following formulas.

$$\begin{aligned} E(X) = \mu_X &= \sum x_i \cdot P(x_i) \\ &= x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + \cdots + x_n \cdot P(x_n) \\ Var(X) = \sigma_x^2 &= \sum (x_i - \mu_X)^2 \cdot P(x_i) \\ SD(X) = \sigma_X &= \sqrt{\sum (x_i - \mu_X)^2 \cdot P(x_i)} \\ &= \sqrt{(x_1 - \mu_X)^2 \cdot P(x_1) + (x_2 - \mu_X)^2 \cdot P(x_2) + \cdots + (x_n - \mu_X)^2 \cdot P(x_n)} \end{aligned}$$

  We can think of $P(x_i)$ as the *weight*, and each term is weighted its appropriate amount.
- The **mean** of a probability distribution does not need to be a value in the distribution. It represents the average of many, many repetitions of a random process. The **standard deviation** represents the typical variation of the outcomes from the mean, when the random process is repeated over and over.
- **Linear transformations**. Adding a constant to every value in a probability distribution adds that value to the mean, but it does not affect the standard deviation. When multiplying every value by a constant, this multiplies the mean by the constant and it multiplies the standard deviation by the absolute value of the constant.
- **Combining random variables**. Let X and Y be random variables and let a and b be constants.
  - The expected value of the sum is the sum of the expected values.
    $E(X + Y) = E(X) + E(Y)$
    $E(aX + bY) = a \times E(X) + b \times E(Y)$
  - When X and Y are **independent**: The standard deviation of a sum or a difference is the square root of the sum of each standard deviation squared.
    $SD(X + Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$
    $SD(X - Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$
    $SD(aX + bY) = \sqrt{(a \times SD(X))^2 + (b \times SD(Y))^2}$

  The SD properties require that X and Y be independent. The expected value properties hold true whether or not X and Y are independent.
- Because the sum or difference of two normally distributed variables is itself a normally distributed variable, the normal approximation is also used in the following type of problem.

  Find the probability that a sum $X + Y$ or a difference $X - Y$ is greater/less than some value.

  1. Verify that the distribution of X and the distribution of Y are approximately normal.
  2. Find the mean of the sum or difference. Recall: the mean of a sum is the sum of the means. The mean of a difference is the difference of the means.
     Find the SD of the sum or difference using:
     $SD(X + Y) = SD(X - Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$.
  3. Calculate the Z-score. Use the calculated mean and SD to standardize the given sum or difference.
  4. Find the appropriate area under the normal curve.

Exercises

3.33 College smokers. At a university, 13% of students smoke.

(a) Calculate the expected number of smokers in a random sample of 100 students from this university.

(b) The university gym opens at 9 am on Saturday mornings. One Saturday morning at 8:55 am there are 27 students outside the gym waiting for it to open. Should you use the same approach from part (a) to calculate the expected number of smokers among these 27 students?

3.34 Ace of clubs wins. Consider the following card game with a well-shuffled deck of cards. If you draw a red card, you win nothing. If you get a spade, you win \$5. For any club, you win \$10 plus an extra \$20 for the ace of clubs.

(a) Create a probability model for the amount you win at this game. Also, find the expected winnings for a single game and the standard deviation of the winnings.

(b) What is the maximum amount you would be willing to pay to play this game? Explain your reasoning.

3.35 Hearts win. In a new card game, you start with a well-shuffled full deck and draw 3 cards without replacement. If you draw 3 hearts, you win \$50. If you draw 3 black cards, you win \$25. For any other draws, you win nothing.

(a) Create a probability model for the amount you win at this game, and find the expected winnings. Also compute the standard deviation of this distribution.

(b) If the game costs \$5 to play, what would be the expected value and standard deviation of the net profit (or loss)? *(Hint: profit = winnings − cost; $X - 5$)*

(c) If the game costs \$5 to play, should you play this game? Explain.

3.36 Is it worth it? Andy is always looking for ways to make money fast. Lately, he has been trying to make money by gambling. Here is the game he is considering playing: The game costs \$2 to play. He draws a card from a deck. If he gets a number card (2-10), he wins nothing. For any face card (jack, queen or king), he wins \$3. For any ace, he wins \$5, and he wins an *extra* \$20 if he draws the ace of clubs.

(a) Create a probability model and find Andy's expected profit per game.

(b) Would you recommend this game to Andy as a good way to make money? Explain.

3.37 Portfolio return. A portfolio's value increases by 18% during a financial boom and by 9% during normal times. It decreases by 12% during a recession. What is the expected return on this portfolio if each scenario is equally likely?

3.38 Baggage fees. An airline charges the following baggage fees: \$25 for the first bag and \$35 for the second. Suppose 54% of passengers have no checked luggage, 34% have one piece of checked luggage and 12% have two pieces. We suppose a negligible portion of people check more than two bags.

(a) Build a probability model, compute the average revenue per passenger, and compute the corresponding standard deviation.

(b) About how much revenue should the airline expect for a flight of 120 passengers? With what standard deviation? Note any assumptions you make and if you think they are justified.

3.39 American roulette. The game of American roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. Gamblers can place bets on red or black. If the ball lands on their color, they double their money. If it lands on another color, they lose their money. Suppose you bet \$1 on red. What's the expected value and standard deviation of your winnings?

3.40 European roulette. The game of European roulette involves spinning a wheel with 37 slots: 18 red, 18 black, and 1 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. Gamblers can place bets on red or black. If the ball lands on their color, they double their money. If it lands on another color, they lose their money.

(a) Suppose you play roulette and bet \$3 on a single round. What is the expected value and standard deviation of your total winnings?

(b) Suppose you bet \$1 in three different rounds. What is the expected value and standard deviation of your total winnings?

(c) How do your answers to parts (a) and (b) compare? What does this say about the riskiness of the two games?

3.41 Lemonade at The Cafe. Drink pitchers at The Cafe are intended to hold about 64 ounces of lemonade and glasses hold about 12 ounces. However, when the pitchers are filled by a server, they do not always fill it with exactly 64 ounces. There is some variability. Similarly, when they pour out some of the lemonade, they do not pour exactly 12 ounces. The amount of lemonade in a pitcher is normally distributed with mean 64 ounces and standard deviation 1.732 ounces. The amount of lemonade in a glass is normally distributed with mean 12 ounces and standard deviation 1 ounce.

(a) How much lemonade would you expect to be left in a pitcher after pouring one glass of lemonade?

(b) What is the standard deviation of the amount left in a pitcher after pouring one glass of lemonade?

(c) What is the probability that more than 50 ounces of lemonade is left in a pitcher after pouring one glass of lemonade?

3.42 Spray paint, Part I. Suppose the area that can be painted using a single can of spray paint is slightly variable and follows a nearly normal distribution with a mean of 25 square feet and a standard deviation of 3 square feet. Suppose also that you buy three cans of spray paint.

(a) How much area would you expect to cover with these three cans of spray paint?

(b) What is the standard deviation of the area you expect to cover with these three cans of spray paint?

(c) The area you wanted to cover is 80 square feet. What is the probability that you will be able to cover this entire area with these three cans of spray paint?

3.43 GRE scores, Part III. In Exercises 2.27 and 2.29 we saw two distributions for GRE scores: $N(\mu = 151, \sigma = 7)$ for the verbal part of the exam and $N(\mu = 153, \sigma = 7.67)$ for the quantitative part. Suppose performance on these two sections is independent. Use this information to compute each of the following:

(a) The probability of a combined (verbal + quantitative) score above 320.

(b) The score of a student who scored better than 90% of the test takers overall.

3.44 Betting on dinner, Part I. Suppose a restaurant is running a promotion where prices of menu items are random following some underlying distribution. If you're lucky, you can get a basket of fries for \$3, or if you're not so lucky you might end up having to pay \$10 for the same menu item. The price of basket of fries is drawn from a normal distribution with mean \$6 and standard deviation of \$2. The price of a fountain drink is drawn from a normal distribution with mean \$3 and standard deviation of \$1. What is the probability that you pay more than \$10 for a dinner consisting of a basket of fries and a fountain drink?

3.5 Geometric distribution

How many times should we expect to roll a die until we get a `1`? How many people should we expect to see at a hospital until we get someone with blood type O+? These questions can be answered using the geometric distribution.

Learning objectives

1. Determine if a scenario is geometric.
2. Calculate the probabilities of the possible values of a geometric random variable.
3. Find and interpret the mean (expected value) and standard deviation of a geometric distribution.
4. Understand the shape of the geometric distribution.

3.5.1 Bernoulli distribution

Many health insurance plans in the United States have a deductible, where the insured individual is responsible for costs up to the deductible, and then the costs above the deductible are shared between the individual and insurance company for the remainder of the year.

Suppose a health insurance company found that 70% of the people they insure stay below their deductible in any given year. Each of these people can be thought of as a **trial**. We label a person a **success** if her healthcare costs do not exceed the deductible. We label a person a **failure** if she does exceed her deductible in the year. Because 70% of the individuals will not exceed their deductible, we denote the **probability of a success** as $p = 0.7$. The probability of a failure is sometimes denoted with $q = 1 - p$, which would be 0.3 in for the insurance example.

When an individual trial only has two possible outcomes, often labeled as `success` or `failure`, it is called a **Bernoulli random variable**. We chose to label a person who does not exceed her deductible as a "success" and all others as "failures". However, we could just as easily have reversed these labels. The mathematical framework we will build does not depend on which outcome is labeled a success and which a failure, as long as we are consistent.

Bernoulli random variables are often denoted as `1` for a success and `0` for a failure. In addition to being convenient in entering data, it is also mathematically handy. Suppose we observe ten trials:

`1 1 1 0 1 0 0 1 1 0`

Then the **sample proportion**, $\hat{p}$, is the sample mean of these observations:

$$\hat{p} = \frac{\text{\# of successes}}{\text{\# of trials}} = \frac{1+1+1+0+1+0+0+1+1+0}{10} = 0.6$$

This mathematical inquiry of Bernoulli random variables can be extended even further. Because `0` and `1` are numerical outcomes, we can define the mean and standard deviation of a Bernoulli random variable.[70]

[70] If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$\begin{aligned}\mu = E(X) &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ &= 0 \cdot (1 - p) + 1 \cdot p = 0 + p = p\end{aligned}$$

Similarly, the variance of X can be computed:

$$\begin{aligned}\sigma^2 &= (0 - p)^2 \cdot P(X = 0) + (1 - p)^2 \cdot P(X = 1) \\ &= p^2 \cdot (1 - p) + (1 - p)^2 \cdot p = p(1 - p)\end{aligned}$$

The standard deviation is $\sigma = \sqrt{p(1-p)}$.

BERNOULLI RANDOM VARIABLE

If X is a random variable that takes value 1 with probability of success p and 0 with probability $1-p$, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p \qquad \sigma = \sqrt{p(1-p)}$$

In general, it is useful to think about a Bernoulli random variable as a random process with only two outcomes: a success or failure. Then we build our mathematical framework using the numerical labels `1` and `0` for successes and failures, respectively.

3.5.2 Geometric distribution

The **geometric distribution** is used to describe how many trials it takes to observe a success. Let's first look at an example.

EXAMPLE 3.83

Suppose we are working at the insurance company and need to find a case where the person did not exceed her (or his) deductible as a case study. If the probability a person will not exceed her deductible is 0.7 and we are drawing people at random, what are the chances that the first person will not have exceeded her deductible, i.e. be a success? The second person? The third? What about the probability that we pull $x-1$ cases before we find the first success, i.e. the first success is the x^{th} person? (If the first success is the fifth person, then we say $x = 5$.)

(E) The probability of stopping after the first person is just the chance the first person will not exceed her (or his) deductible: 0.7. The probability the second person is the first to exceed her deductible is

$$\begin{aligned} &P(\text{second person is the first to exceed deductible}) \\ &\quad = P(\text{the first won't, the second will}) = (0.3)(0.7) = 0.21 \end{aligned}$$

Likewise, the probability it will be the third case is $(0.3)(0.3)(0.7) = 0.063$.

If the first success is on the x^{th} person, then there are $x-1$ failures and finally 1 success, which corresponds to the probability $(0.3)^{x-1}(0.7)$. This is the same as $(1-0.7)^{x-1}(0.7)$.

Example 3.83 illustrates what the **geometric distribution**, which describes the waiting time until a success for **independent and identically distributed (iid)** Bernoulli random variables. In this case, the *independence* aspect just means the individuals in the example don't affect each other, and *identical* means they each have the same probability of success.

The geometric distribution from Example 3.83 is shown in Figure 3.27. In general, the probabilities for a geometric distribution decrease **exponentially** fast.

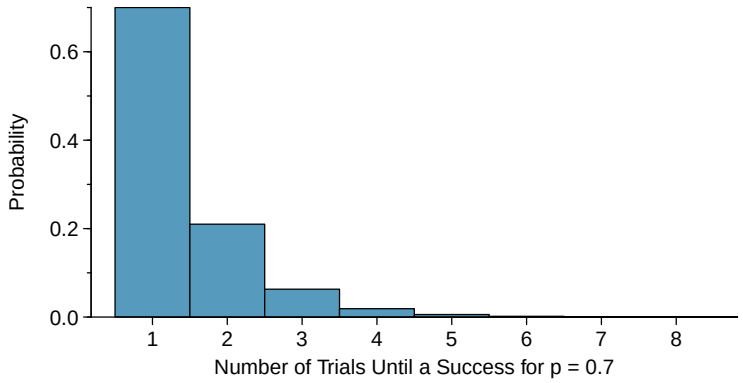


Figure 3.27: The geometric distribution when the probability of success is $p = 0.7$.

While this text will not derive the formulas for the mean (expected) number of trials needed to find the first success or the standard deviation of this distribution, we present general formulas for each.

GEOMETRIC DISTRIBUTION

Let X have a geometric distribution with one parameter p, where p is the probability of a success in one trial. Then the probability of finding the first success in the x^{th} trial is given by

$$P(X = x) = (1 - p)^{x-1}p$$

where $x = 1, 2, 3, \ldots$

The mean (i.e. expected value) and standard deviation of this wait time are given by

$$\mu_X = \frac{1}{p} \qquad \sigma_X = \frac{\sqrt{1-p}}{p}$$

It is no accident that we use the symbol μ for both the mean and expected value. The mean and the expected value are one and the same.

It takes, on average, $1/p$ trials to get a success under the geometric distribution. This mathematical result is consistent with what we would expect intuitively. If the probability of a success is high (e.g. 0.8), then we don't usually wait very long for a success: $1/0.8 = 1.25$ trials on average. If the probability of a success is low (e.g. 0.1), then we would expect to view many trials before we see a success: $1/0.1 = 10$ trials.

GUIDED PRACTICE 3.84

(G) The probability that a particular case would not exceed their deductible is said to be 0.7. If we were to examine cases until we found one that where the person did not exceed her deductible, how many cases should we expect to check?[71]

EXAMPLE 3.85

What is the chance that we would find the first success within the first 3 cases?

This is the chance the first ($X = 1$), second ($X = 2$), or third ($X = 3$) case is the first success, which are three disjoint outcomes. Because the individuals in the sample are randomly sampled from a large population, they are independent. We compute the probability of each case and add the separate results:

(E)

$$\begin{aligned} &P(X = 1, 2, \text{ or } 3) \\ &\quad = P(X = 1) + P(X = 2) + P(X = 3) \\ &\quad = (0.3)^{1-1}(0.7) + (0.3)^{2-1}(0.7) + (0.3)^{3-1}(0.7) \\ &\quad = 0.973 \end{aligned}$$

There is a probability of 0.973 that we would find a successful case within 3 cases.

GUIDED PRACTICE 3.86

(G) Determine a more clever way to solve Example 3.85. Show that you get the same result.[72]

[71] We would expect to see about $1/0.7 \approx 1.43$ individuals to find the first success.

[72] First find the probability of the complement: $P(\text{no success in first 3 trials}) = 0.3^3 = 0.027$. Next, compute one minus this probability: $1 - P(\text{no success in 3 trials}) = 1 - 0.027 = 0.973$.

EXAMPLE 3.87

Suppose a car insurer has determined that 88% of its drivers will not exceed their deductible in a given year. If someone at the company were to randomly draw driver files until they found one that had not exceeded their deductible, what is the expected number of drivers the insurance employee must check? What is the standard deviation of the number of driver files that must be drawn?

E

In this example, a success is again when someone will not exceed the insurance deductible, which has probability $p = 0.88$. The expected number of people to be checked is $1/p = 1/0.88 = 1.14$ and the standard deviation is $\frac{\sqrt{1-p}}{p} = \frac{\sqrt{1-0.88}}{0.88} = 0.39$.

[reworded]

GUIDED PRACTICE 3.88

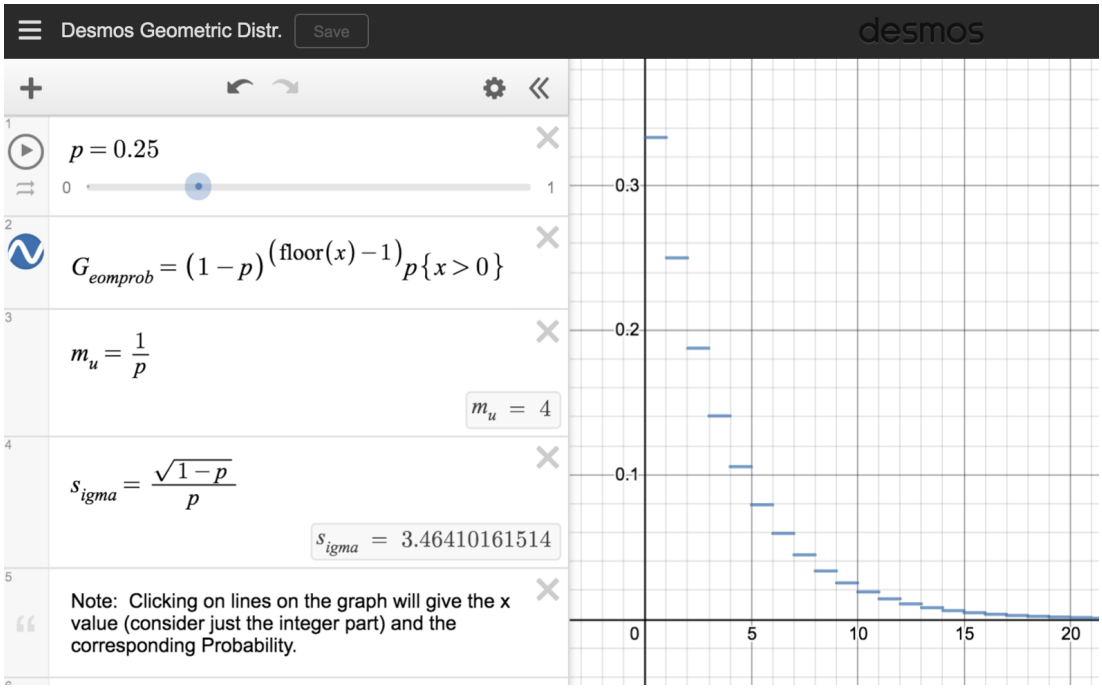
Using the results from Example 3.87, $\mu_X = 1.14$ and $\sigma_X = 0.39$, would it be appropriate to use the empirical rule to find what proportion of experiments would end in 1.14 ± 0.39 trials?[73]

G

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 x^{th} trial, we had to use the General Multiplication Rule for independent processes. It is no simple task to generalize the geometric model for dependent trials.

3.5.3 Technology: geometric probablities

Get started quickly with this Desmos Geometric Calculator.



[73] No. The geometric distribution is always right skewed and can never be well-approximated by a normal model.

Section summary

- It is useful to model yes/no, success/failure with the values 1 and 0, respectively. We call the **probability of success** p and the **probability of failure** $1-p$.
- When the trials are **independent** and the value of p is constant, the probability of finding **the first success on the x^{th} trial** is given by $(1-p)^{x-1}p$. We can see the reasoning behind this formula as follows: for the first success to happen on the x^{th} trial, it has to *not* happen the first $x-1$ trials (with probability $1-p$), and then happen on the x^{th} trial (with probability p).
- When we consider the *entire distribution* of possible values for the how long *until* the first success, we get a discrete probability distribution known as the geometric distribution. The **geometric distribution** describes the waiting time *until* the first success, when the trials are independent and the probability of success, p, is constant. If X has a geometric distribution with parameter p, then $P(X=x)=(1-p)^{x-1}p$, where $x=1,2,3\ldots$.
- The geometric distribution is always *right skewed* and, in fact, has no maximum value. The probabilities, though, decrease exponentially fast.
- Even though the geometric distribution has an infinite number of values, it has a well-defined **mean**: $\mu_X=\frac{1}{p}$ and **standard deviation**: $\sigma_X=\frac{\sqrt{1-p}}{p}$. If the probability of success is $\frac{1}{10}$, then *on average* it takes 10 trials until we see the first success.
- Note that when the trials are not independent, we can modify the geometric formula to find the probability that the first success happens on the x^{th} trial. Instead of simply raising $(1-p)$ to the $x-1$, multiply the appropriate *conditional* probabilities.

Exercises

3.45 Is it Bernoulli? Determine if each trial can be considered an independent Bernoulli trial for the following situations.

(a) Cards dealt in a hand of poker.

(b) Outcome of each roll of a die.

3.46 With and without replacement. In the following situations assume that half of the specified population is male and the other half is female.

(a) Suppose you're sampling from a room with 10 people. What is the probability of sampling two females in a row when sampling with replacement? What is the probability when sampling without replacement?

(b) Now suppose you're sampling from a stadium with 10,000 people. What is the probability of sampling two females in a row when sampling with replacement? What is the probability when sampling without replacement?

(c) We often treat individuals who are sampled from a large population as independent. Using your findings from parts (a) and (b), explain whether or not this assumption is reasonable.

3.47 Eye color, Part I. A husband and wife both have brown eyes but carry genes that make it possible for their children to have brown eyes (probability 0.75), blue eyes (0.125), or green eyes (0.125).

(a) What is the probability the first blue-eyed child they have is their third child? Assume that the eye colors of the children are independent of each other.

(b) On average, how many children would such a pair of parents have before having a blue-eyed child? What is the standard deviation of the number of children they would expect to have until the first blue-eyed child?

3.48 Defective rate. A machine that produces a special type of transistor (a component of computers) has a 2% defective rate. The production is considered a random process where each transistor is independent of the others.

(a) What is the probability that the 10^{th} transistor produced is the first with a defect?

(b) What is the probability that the machine produces no defective transistors in a batch of 100?

(c) On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?

(d) Another machine that also produces transistors has a 5% defective rate where each transistor is produced independent of the others. On average how many transistors would you expect to be produced with this machine before the first with a defect? What is the standard deviation?

(e) Based on your answers to parts (c) and (d), how does increasing the probability of an event affect the mean and standard deviation of the wait time until success?

3.49 Bernoulli, the mean. Use the probability rules from Section 3.4 to derive the mean of a Bernoulli random variable, i.e. a random variable X that takes value 1 with probability p and value 0 with probability $1 - p$. That is, compute the expected value of a generic Bernoulli random variable.

3.50 Bernoulli, the standard deviation. Use the probability rules from Section 3.4 to derive the standard deviation of a Bernoulli random variable, i.e. a random variable X that takes value 1 with probability p and value 0 with probability $1 - p$. That is, compute the square root of the variance of a generic Bernoulli random variable.

3.6 Binomial distribution

What is the probability of exactly 50 heads in 100 coin tosses? Or the probability of randomly sampling 12 people and having more than 9 of them identify as male? If the probability of a defective part is 1%, how many defective items would we expect in a random shipment of 200 of those parts? We can model these scenarios and answer these questions using the binomial distribution.

Learning objectives

1. Calculate the number of possible scenarios for obtaining x successes in n trials.
2. Determine whether a scenario is binomial or not.
3. Calculate the probabilities of the possible values of a binomial random variable using the binomial formula.
4. Recognize that the binomial formula uses the special Addition Rule for mutually exclusive events.
5. Find probabilities of the form "at least or "at most by applying the binomial formula multiple times.
6. Calculate and interpret the mean (expected value) and standard deviation of the number of successes in n binomial trials.
7. Determine whether a binomial distribution can be modeled as approximately normal. If so, use normal approximation to estimate cumulative binomial probabilities.

3.6.1 Introducing the binomial formula

Let's again imagine ourselves back at the insurance agency where 70% of individuals do not exceed their deductible. Each people is thought of as a **trial**. We label a trial a **success** if the individual healthcare costs do not exceed the deductible. We label a trial a **failure** if the individual healthcare costs do exceed the deductible. Because 70% of the individuals will not exceed their deductible, we denote the **probability of a success** as $p = 0.7$.

EXAMPLE 3.89

Suppose the insurance agency is considering a random sample of four individuals they insure. What is the chance exactly one of them will exceed the deductible and the other three will not? Let's call the four people Ariana (A), Brittany (B), Carlton (C), and Damian (D) for convenience.

Let's consider a scenario where one person exceeds the deductible:

(E)

$$\begin{aligned} &P(A = \texttt{exceed},\ B = \texttt{not},\ C = \texttt{not},\ D = \texttt{not}) \\ &\quad = P(A = \texttt{exceed})\ P(B = \texttt{not})\ P(C = \texttt{not})\ P(D = \texttt{not}) \\ &\quad = (0.3)(0.7)(0.7)(0.7) \\ &\quad = (0.7)^3(0.3)^1 \\ &\quad = 0.103 \end{aligned}$$

But there are three other scenarios: Brittany, Carlton, or Damian could have been the one to exceed the deductible. In each of these cases, the probability is again $(0.7)^3(0.3)^1$. These four scenarios exhaust all the possible ways that exactly one of these four people could have exceeded the deductible, so the total probability is $4 \times (0.7)^3(0.3)^1 = 0.412$.

GUIDED PRACTICE 3.90

Verify that the scenario where Brittany is the only one to exceed the deductible has probability $(0.7)^3(0.3)^1$. [74]

The binomial distribution describes the probability of having exactly x successes in n independent trials with probability of a success p (in Example 3.89, $n = 4$, $x = 3$, $p = 0.7$). We would like to determine the probabilities associated with the binomial distribution more generally, i.e. we want a formula where we can use n, x, and p to obtain the probability. To do this, we reexamine each part of Example 3.89.

There were four individuals who could have been the one to exceed the deductible, and each of these four scenarios had the same probability. Thus, we could identify the final probability as

$$[\#\text{ of scenarios}] \times P(\text{single scenario})$$

The first component of this equation is the number of ways to arrange the $x = 3$ successes among the $n = 4$ trials. The second component is the probability of any of the four (equally probable) scenarios.

Consider $P(\text{single scenario})$ under the general case of x successes and $n - x$ failures in the n trials. In any such scenario, we apply the Multiplication Rule for independent events:

$$p^x(1-p)^{n-x}$$

This is our general formula for $P(\text{single scenario})$.

Secondly, we introduce the **binomial coefficient**, which gives the number of ways to choose x successes in n trials, i.e. arrange x successes and $n - x$ failures:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

The quantity $\binom{n}{x}$ is read **n choose x**.[75] The exclamation point notation (e.g. $n!$) denotes a **factorial** expression.

$$\begin{aligned}
0! &= 1 \\
1! &= 1 \\
2! &= 2 \times 1 = 2 \\
3! &= 3 \times 2 \times 1 = 6 \\
4! &= 4 \times 3 \times 2 \times 1 = 24 \\
&\vdots \\
n! &= n \times (n-1) \times ... \times 3 \times 2 \times 1
\end{aligned}$$

Using the formula, we can compute the number of ways to choose $x = 3$ successes in $n = 4$ trials:

$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4!}{3!1!} = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(1)} = 4$$

This result is exactly what we found by carefully thinking of each possible scenario in Example 3.89.

Substituting n choose x for the number of scenarios and $p^x(1-p)^{n-x}$ for the single scenario probability yields the **binomial formula**.

[74] $P(A = \texttt{not},\ B = \texttt{exceed},\ C = \texttt{not},\ D = \texttt{not}) = (0.7)(0.3)(0.7)(0.7) = (0.7)^3(0.3)^1$.

[75] Other notations for n choose x includes ${}_nC_x$, C_n^x, and $C(n, x)$.

BINOMIAL FORMULA

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly x successes in n independent trials is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

3.6.2 When and how to apply the formula

IS IT BINOMIAL? FOUR CONDITIONS TO CHECK.

(1) The trials are independent.
(2) The number of trials, n, is fixed.
(3) Each trial outcome can be classified as a *success* or *failure*.
(4) The probability of a success, p, is the same for each trial.

EXAMPLE 3.91

What is the probability that 3 of 8 randomly selected individuals will have exceeded the insurance deductible, i.e. that 5 of 8 will not exceed the deductible? Recall that 70% of individuals will not exceed the deductible.

We would like to apply the binomial model, so we check the conditions. The number of trials is fixed ($n = 8$) (condition 2) and each trial outcome can be classified as a success or failure (condition 3). Because the sample is random, the trials are independent (condition 1) and the probability of a success is the same for each trial (condition 4).

In the outcome of interest, there are $x = 5$ successes in $n = 8$ trials (recall that a success is an individual who does *not* exceed the deductible, and the probability of a success is $p = 0.7$. So the probability that 5 of 8 will not exceed the deductible and 3 will exceed the deductible is given by

(E)

$$\begin{aligned}\binom{8}{5}(0.7)^5(1-0.7)^{8-5} &= \frac{8!}{5!(5-3)!}(0.7)^5(1-0.7)^{8-5} \\ &= \frac{8!}{5!3!}(0.7)^5(0.3)^3\end{aligned}$$

Dealing with the factorial part:

$$\frac{8!}{5!3!} = \frac{8\times7\times6\times5\times4\times3\times2\times1}{(5\times4\times3\times2\times1)(3\times2\times1)} = \frac{8\times7\times6}{3\times2\times1} = 56$$

Using $(0.7)^5(0.3)^3 \approx 0.00454$, the final probability is about $56 \times 0.00454 \approx 0.254$.

If you must calculate the binomial coefficient by hand, it's often useful to cancel out as many terms as possible in the top and bottom. See Section 3.6.3 for how to evaluate the binomial coefficient and the binomial formula using a calculator.

COMPUTING BINOMIAL PROBABILITIES

The first step in using the binomial model is to check that the model is appropriate. The second step is to identify n, p, and x. Finally, apply the binomial formula to determine the probability and interpret the results.

[added without replacement consideration]

EXAMPLE 3.92

Approximately 35% of a population has blood type O+. Suppose four people show up at a hospital and we want to find the probability that exactly one of them has blood type O+. Can we use the binomial formula?

To check if the binomial model is appropriate, we must verify the conditions.

E

1. We will suppose that these 4 people comprise a random sample. This seems reasonable, since one person with a particular blood type showing up at a hospital seems unlikely to affect the chance that other people with that blood type would show up at the hospital. This sample is without replacement, though, not with replacement, so the observations are not entirely independent. However, when the sample size is very small compared to the population size, we can treat the observations *as if they were with replacement*, since the composition of the population changes very little with each additional person sampled. Since we have a random sample of a very small percent of the population, we will consider the independence condition met.
2. We have a fixed number of trials ($n = 4$).
3. Each outcome is a success or failure (blood type O+ or not blood type O+).
4. The probability of a success is the same for each trial since the individuals are like a random sample ($p = 0.35$ if we say a "success" is someone having blood type O+).

SAMPLING WITHOUT REPLACEMENT

When randomly sampling without replacement, if the sample size is small relative to the population size (rule of thumb: sample size less than 1/10 of the population size), we will consider the observations to be independent.

EXAMPLE 3.93

Given that 35% of a population has blood type O+, what is the probabilty that in a random sample of 4 people:

(a) none of them have blood type O+?

(b) one will have blood type O+?

(c) no more than one will have blood type O+?

E

(a) $P(X = 0) = \binom{4}{0}(0.35)^0(0.65)^4 = 1 \times 1 \times 0.65^4 = 0.65^4 = 0.179$
Note that we could have answered this question without the binomial formula, using methods from the previous section.

(b) $P(X = 1) = \binom{4}{1}(0.35)^1(0.65)^3 = 0.384$.

(c) This can be computed as the sum of parts (a) and (b): $P(X = 0) + P(X = 1) = 0.179 + 0.384 = 0.563$. That is, there is about a 56.3% chance that no more than one of them will have blood type O+.

GUIDED PRACTICE 3.94

G

What is the probability that at least 3 of 4 people in a random sample will have blood type O+ if 35% of the population has blood type O+?[76]

[76] $P(\text{at least 3 of 4 have blood type O+}) = P(X = 3) + P(X = 4) = \binom{4}{3}(0.35)^3(0.65)^1 + (0.35)^4 = 0.111 + 0.015 = 0.126$

GUIDED PRACTICE 3.95

The probability that a random smoker will develop a severe lung condition in his or her lifetime is about 0.3. If you have 4 friends who smoke and you want to find the probability that 1 of them will develop a severe lung condition in his or her lifetime, can you apply the binomial formula?[77]

EXAMPLE 3.96

There are 13 marbles in a bag. 4 are blue and 9 are red. Randomly draw 5 marbles *without replacement*. Find the probability you get exactly 3 blue marbles. Because we are drawing without replacement the probability of success p is not the same for each trial. Also, the sample size is large compared to the population size (much greater than 1/10 of the population size), so we cannot treat these observations as independent and we cannot use the binomial formula. However, we can use the same logic to arrive at the following answer.

$$\begin{aligned} P(X=3) &= (\#\text{ of combinations with 3 blue}) \times P(\text{3 blue and 2 red in a } \textit{specific} \text{ order}) \\ &= \binom{5}{3} \times P(\text{BBBRR}) \\ &= \binom{5}{3}\left(\frac{4}{13}\times\frac{3}{12}\times\frac{2}{11}\times\frac{9}{10}\times\frac{8}{9}\right) \\ &= 0.1119 \end{aligned}$$

GUIDED PRACTICE 3.97

Draw 4 cards without replacement from a deck of 52 cards. What is the probability that you get at least two hearts?[78]

Lastly, we consider the binomial coefficient,, n choose x, under some special scenarios.

GUIDED PRACTICE 3.98

Why is it true that $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ for any number n?[79]

GUIDED PRACTICE 3.99

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?[80]

---

[77] While conditions (2) and (3) are met, most likely the friends know each other, so the independence assumption (1) is probably not satisfied. For example, acquaintances may have similar smoking habits, or those friends might make a pact to quit together. Condition (4) is also not satisfied since this is not a random sample of people.

[78] $P(\text{at least 2 hearts in 4 draws from a deck}) = 1 - [P(X=0) + P(X=1)] = 1 - [(\frac{39}{52})(\frac{38}{51})(\frac{37}{50})(\frac{36}{49}) + \binom{4}{1}(\frac{13}{52})(\frac{39}{51})(\frac{38}{50})(\frac{37}{49})] = 1 - [.0.3038 + 0.4388] = 0.2574$.

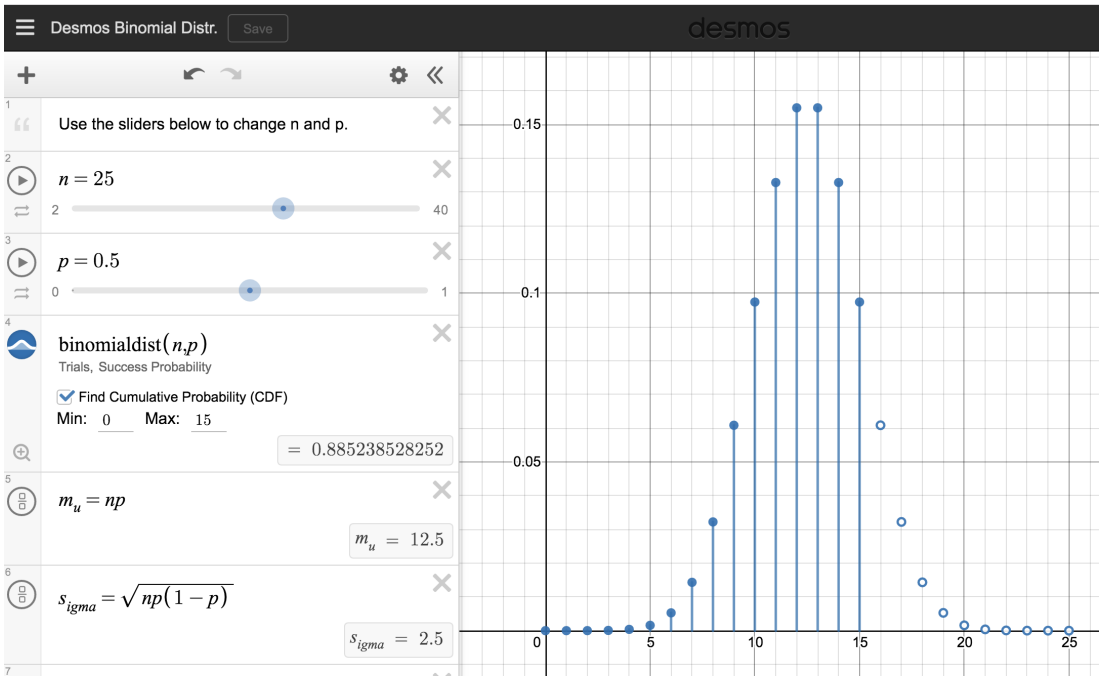
[79] Frame these expressions into words. How many different ways are there to arrange 0 successes and n failures in n trials? (1 way.) How many different ways are there to arrange n successes and 0 failures in n trials? (1 way.)

[80] One success and $n-1$ failures: there are exactly n unique places we can put the success, so there are n ways to arrange one success and $n-1$ failures. A similar argument is used for the second question. Mathematically, we show these results by verifying the following two equations:

$$\binom{n}{1} = n, \qquad \binom{n}{n-1} = n$$

3.6.3 Technology: binomial probabilities

Get started quickly with this Desmos Binomial Calculator.



Calculator instructions

TI-83/84: COMPUTING THE BINOMIAL COEFFICIENT $\binom{n}{x}$

Use `MATH`, `PRB`, `nCr` to evaluate n choose r. Here r and x are different letters for the same quantity.

1. Type the value of n.
2. Select `MATH`.
3. Right arrow to `PRB`.
4. Choose `3:nCr`.
5. Type the value of x.
6. Hit `ENTER`.

Example: `5 nCr 3` means 5 choose 3.

CASIO FX-9750GII: COMPUTING THE BINOMIAL COEFFICIENT $\binom{n}{x}$

1. Navigate to the `RUN-MAT` section (hit `MENU`, then hit `1`).
2. Enter a value for n.
3. Go to `CATALOG` (hit buttons `SHIFT` and then `7`).
4. Type `C` (hit the `ln` button), then navigate down to the bolded `C` and hit `EXE`.
5. Enter the value of x. Example of what it should look like: `7C3`.
6. Hit `EXE`.

TI-84: COMPUTING THE BINOMIAL FORMULA, $P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$

Use 2ND VARS, binompdf to evaluate the probability of *exactly* x occurrences out of n independent trials of an event with probability p.

1. Select 2ND VARS (i.e. DISTR)
2. Choose A:binompdf (use the down arrow to scroll down).
3. Let trials be n.
4. Let p be p
5. Let x value be x.
6. Select Paste and hit ENTER.

TI-83: Do step 1, choose 0:binompdf, then enter n, p, and x separated by commas: binompdf(n, p, x). Then hit ENTER.

TI-84: COMPUTING $P(X \le x) = \binom{n}{0}p^0(1-p)^{n-0} + ... + \binom{n}{x}p^x(1-p)^{n-x}$

Use 2ND VARS, binomcdf to evaluate the cumulative probability of *at most* x occurrences out of n independent trials of an event with probability p.

1. Select 2ND VARS (i.e. DISTR)
2. Choose B:binomcdf (use the down arrow).
3. Let trials be n.
4. Let p be p
5. Let x value be x.
6. Select Paste and hit ENTER.

TI-83: Do steps 1-2, then enter the values for n, p, and x separated by commas as follows: binomcdf(n, p, x). Then hit ENTER.

CASIO FX-9750GII: BINOMIAL CALCULATIONS

1. Navigate to STAT (MENU, then hit 2).
2. Select DIST (F5), and then BINM (F5).
3. Choose whether to calculate the binomial distribution for a specific number of successes, $P(X = k)$, or for a range $P(X \le k)$ of values (0 successes, 1 success, ..., x successes).
   - For a specific number of successes, choose Bpd (F1).
   - To consider the range 0, 1, ..., x successes, choose Bcd(F1).
4. If needed, set Data to Variable (Var option, which is F2).
5. Enter the value for x (x), Numtrial (n), and p (probability of a success).
6. Hit EXE.

GUIDED PRACTICE 3.100

Find the number of ways of arranging 3 blue marbles and 2 red marbles.[81]

[81] Here $n = 5$ and $x = 3$. Doing 5 nCr 3 gives the number of combinations as 10.

GUIDED PRACTICE 3.101

There are 13 marbles in a bag. 4 are blue and 9 are red. Randomly draw 5 marbles *with replacement*. Find the probability you get exactly 3 blue marbles.[82]

GUIDED PRACTICE 3.102

There are 13 marbles in a bag. 4 are blue and 9 are red. Randomly draw 5 marbles *with replacement*. Find the probability you get *at most* 3 blue marbles (i.e. less than or equal to 3 blue marbles).[83]

3.6.4 An example of a binomial distribution

In Guided Practice 3.93, we asked various probability questions regarding the number of people out of 4 with blood type O+. We verified that the scenario was binomial and that each problem could be solved using the binomial formula. Instead of looking at it piecewise, we could describe the entire *distribution* of possible values and their corresponding probabilities. Since there are 4 people, there are several possible outcomes for the number who might have blood type O+: 0, 1, 2, 3, 4. We can make a distribution table with these outcomes. Recall that the probability of a randomly sampled person being blood type O+ is about 0.35.

The **binomial distribution** is used to describe the number of successes in a fixed number of trials. This is different from the geometric distribution, which described the number of trials we must wait before we observe a success.

x_i	$P(x_i)$
0	$\binom{4}{0}(0.35)^0(0.65)^4 = 0.179$
1	$\binom{4}{1}(0.35)^1(0.65)^3 = 0.384$
2	$\binom{4}{2}(0.35)^2(0.65)^2 = 0.311$
3	$\binom{4}{3}(0.35)^3(0.65)^1 = 0.111$
4	$\binom{4}{4}(0.35)^4(0.65)^0 = 0.015$

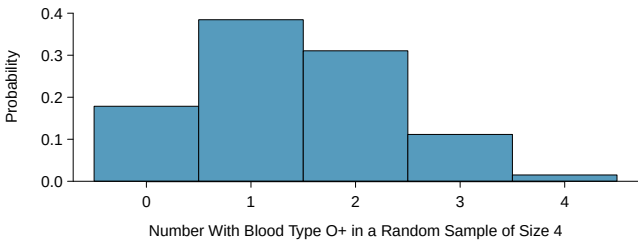


Figure 3.28: Probability distribution for the number with blood type O+ in a random sample of 4 people. This is a binomial distribution. Correcting for rounding error, the probabilities add up to 1, as they must for any probability distribution.

3.6.5 The mean and standard deviation of a binomial distribution

Since this is a probability distribution we could find its mean and standard deviation using the formulas from Chapter 3. Those formulas require a lot of calculations, so it is fortunate there's an easier way to compute the mean and standard deviation for a binomial random variable.

MEAN AND STANDARD DEVIATION OF THE BINOMIAL DISTRIBUTION

For a binomial distribution with parameters n and p, where n is the number of trials and p is the probability of a success, the mean and standard deviation of the number of observed successes are

$$\mu_x = np \qquad\qquad \sigma_x = \sqrt{np(1-p)}$$

[82] Here, $n = 5$, $p = 4/13$, and $x = 3$, so set `trials` $= 5$, `p` $= 4/13$ and `x value` $= 3$. The probability is 0.1396.
[83] Similarly, set `trials` $= 5$, `p` $= 4/13$ and `x value` $= 3$. The cumulative probability is 0.9662.

EXAMPLE 3.103

If the probability that a person has blood type O+ is 0.35 and you have 40 randomly selected people, about how many would you expect to have blood type O+? What is the standard deviation of the number of people who would have blood type O+ among the 40 people?

(E) We are asked to determine the expected number (the mean) and the standard deviation, both of which can be directly computed from the formulas above.

$$\mu_X = np = 40(0.35) = 14$$
$$\sigma_X = \sqrt{np(1-p)} = \sqrt{40(0.35)(0.65)} = 3.0$$

The exact distribution is shown in Figure 3.29.

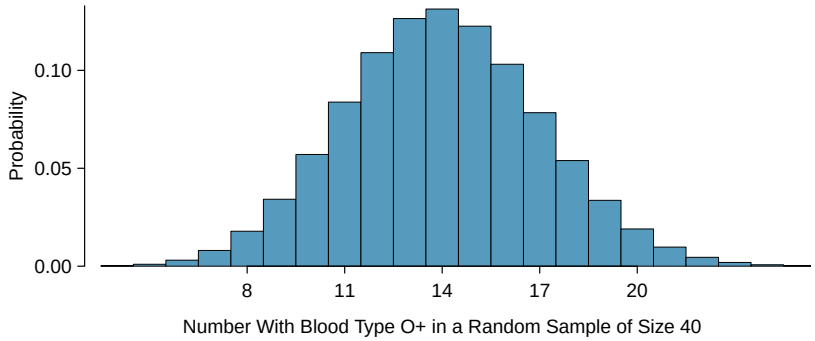


Figure 3.29: Distribution for the number of people with blood type O+ in a random sample of size 40, where $p = 0.35$. The distribution is binomial and is centered on 14 with a standard deviation of 3.

3.6.6 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.

EXAMPLE 3.104

Find the probability that fewer than 12 out of 40 randomly selected people would have blood type O+, where probability of blood type O+ is 0.35.

(E) This is equivalent to asking, what is the probability of observing $X = 0, 1, 2, ...,$ or 11 with blood type O+ in a sample of size 40 when $p = 0.35$? We previously verified that this scenario is binomial. We can compute each of the 12 probabilities using the binomial formula and add them together to find the answer:

$$\begin{aligned} &P(X = 0 \text{ or } X = 1 \text{ or } \cdots \text{ or } X = 11) \\ &\quad = P(X = 0) + P(X = 1) + \cdots + P(X = 11) \\ &\quad = \binom{40}{0}(0.35)^0(0.65)^{40} + \binom{40}{1}(0.35)^1(0.65)^{39} + \cdots + \binom{40}{11}(0.35)^{11}(0.65)^{29} \\ &\quad = 0.21 \end{aligned}$$

If the true proportion with blood type O+ in the population is $p = 0.35$, then the probability of observing fewer than 12 in a sample of $n = 40$ is 0.21.

The computations in Example 3.6.6 are tedious and long. In general, we should avoid such work if an alternative method exists that is faster, easier, and still accurate. Recall that calculating probabilities of a range of values is much easier in the normal model. In some cases we may use the normal distribution to estimate binomial probabilities. While a normal approximation for the distribution in Figure 3.28 when the sample size was $n = 4$ would not be appropriate, it might not be too bad for the distribution in Figure 3.29 where $n = 40$. We might wonder, when is it reasonable to use the normal model to approximate a binomial distribution?

GUIDED PRACTICE 3.105

G Here we consider the binomial model when the probability of a success is $p = 0.10$. Figure 3.30 shows four hollow histograms for simulated samples from the binomial distribution using four different sample sizes: $n = 10$, 30, 100, 300. What happens to the shape of the distributions as the sample size increases? How does the binomial distribution change as n gets larger?[84]

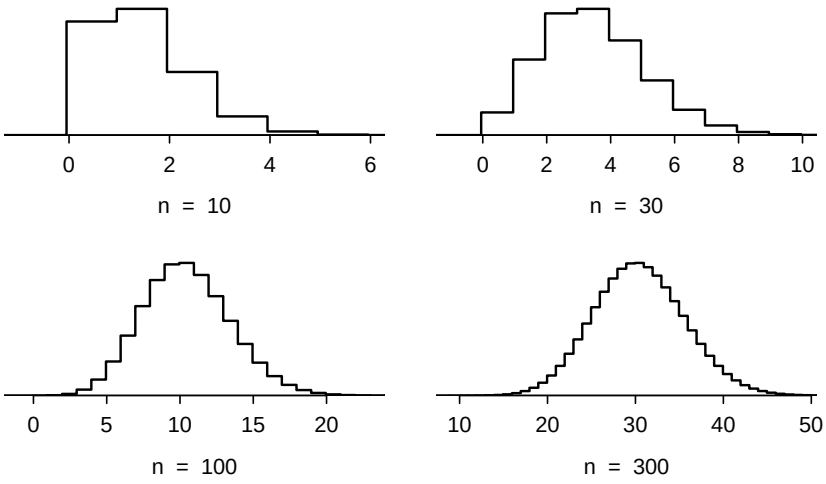


Figure 3.30: Hollow histograms of samples from the binomial model when $p = 0.10$. The sample sizes for the four plots are $n = 10$, 30, 100, and 300, respectively.

The shape of the binomial distribution depends upon both n and p. Here we introduce a rule of thumb for when normal approximation of a binomial distribution is reasonable. We will use this rule of thumb in many applications going forward.

NORMAL APPROXIMATION OF THE BINOMIAL DISTRIBUTION

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that $np \geq 10$ and $n(1-p) \geq 10$. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

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 described in Figure 3.29.

[84] The distribution is transformed from a blocky and skewed distribution into one that rather resembles the normal distribution in the last hollow histogram.

EXAMPLE 3.106

Use the normal approximation to estimate the probability of observing fewer than 12 with blood type O+ in a random sample of 40, if the true proportion with blood type O+ in the population is $p = 0.35$.

First we verify that np and $n(1-p)$ are at least 10 so that we can apply the normal approximation to the binomial model:

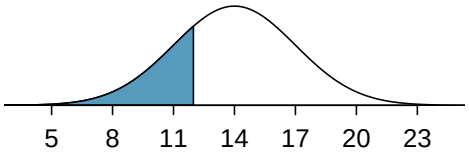
$$np = 40(0.35) = 14 \geq 10 \qquad n(1-p) = 40(0.65) = 26 \geq 10$$

With these conditions checked, we may use the normal distribution to approximate the binomial distribution with the following mean and standard deviation:

(E)

$$\mu = np = 40(0.35) = 14$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{40(0.35)(0.65)} = 3.0$$

We want to find the probability of observing fewer than 12 with blood type O+ using this model. We note that 12 is less than 1 standard deviation below the mean:



Next, we compute the Z-score as $Z = \frac{12-14}{3} = -0.67$ to find the shaded area in the picture: $P(Z < -0.67) = 0.25$. This probability of 0.25 using the normal approximation is reasonably close to the true probability of 0.21 computed using the binomial distribution.

EXAMPLE 3.107

Use the normal approximation to estimate the probability of observing fewer than 120 people with blood type O+ in a random sample of 400, if the true proportion with blood type O+ in the population is $p = 0.35$.

We have previously verified that the binomial model is reasonable for this context. Now we will verify that both np and $n(1-p)$ are at least 10 so we can apply the normal approximation to the binomial model:

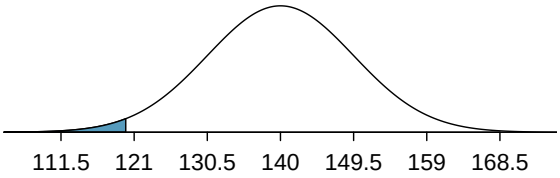
$$np = 400(0.35) = 140 \geq 10 \qquad n(1-p) = 400(0.35) = 260 \geq 10$$

With these conditions checked, we may use the normal approximation in place of the binomial distribution with the following mean and standard deviation:

(E)

$$\mu = np = 400(0.35) = 140$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{400(0.35)(0.65)} = 9.5$$

We want to find the probability of observing fewer than 120 with blood type O+ using this model. We note that 120 is just over 2 standard deviations below the mean:



Next, we compute the Z-score as $Z = \frac{120-140}{9.5} = -2.1$ to find the shaded area in the picture: $P(Z < -2.1) = 0.0179$. This probability of 0.0179 using the normal approximation is very close to the true probability of 0.0196 from the binomial distribution.

GUIDED PRACTICE 3.108

(G) Use normal approximation, if applicable, to estimate the probability of getting greater than 15 sixes in 100 rolls of a fair die.[85]

3.6.7 Normal approximation breaks down on small intervals (special topic)

THE NORMAL APPROXIMATION MAY FAIL ON SMALL INTERVALS

The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

We consider again our example where 35% of people are blood type O+. Suppose we want to find the probability that between 129 and 131 people, inclusive, have blood type O+ in a random sample of 400 people. We want to compute the probability of observing 129, 130, or 131 people with blood type O+ when $p = 0.20$ and $n = 400$. With such a large sample, we might be tempted to apply the normal approximation and use the range 129 to 131. However, we would find that the

[85] $np = 100(1/6) = 16.7 \geq 10$ and $n(1-p) = 100(5/6) = 83.3 \geq 10$
$\mu = np = 100(1/6) = 16.7; \sigma = \sqrt{np(1-p)} = \sqrt{100(1/6)(5/6)} = 3.7$
$Z = \frac{15-16.7}{3.7} = -0.46.$
$P(Z > -0.46) = 0.677$

binomial solution and the normal approximation notably differ:

Binomial: 0.0732 Normal: 0.0483

We can identify the cause of this discrepancy using Figure 3.31, which shows the areas representing the binomial probability (outlined) and normal approximation (shaded). Notice that the width of the area under the normal distribution is 0.5 units too slim on both sides of the interval. The binomial distribution is a discrete distribution, and the each bar is centered over an integer value. Looking closely at Figure 3.31, we can see that the bar corresponding to 129 begins at 128.5 and ends at 129.5, the bar corresponding to 131 begins at 130.5 and ends at 131.5, etc.

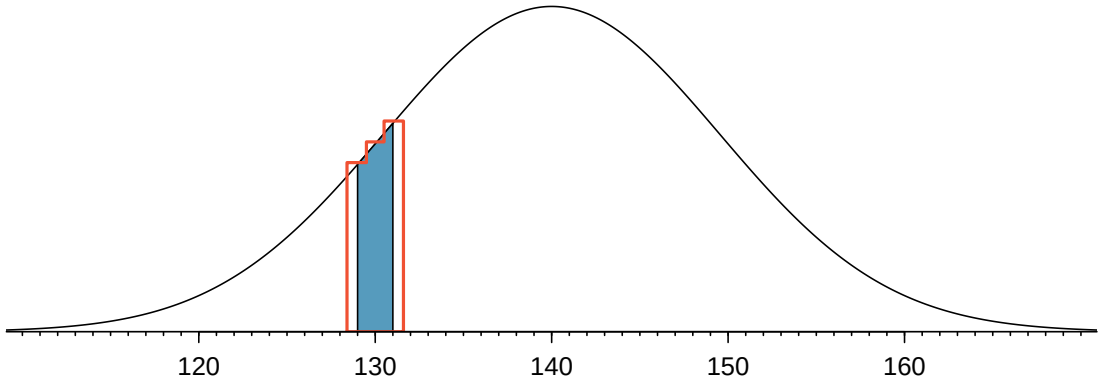


Figure 3.31: A normal curve with the area between 129 and 131 shaded. The outlined area from 128.5 to 131.5 represents the exact binomial probability.

IMPROVING ACCURACY OF THE NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

The normal approximation to the binomial distribution for intervals of values is usually improved if cutoff values for the lower end of a shaded region are reduced by 0.5 and the cutoff value for the upper end are increased by 0.5. This correction is called the continuity correction and accounts for the fact that the binomial distribution is discrete.

EXAMPLE 3.109

Use the method described to find a more accurate estimate for the probability of observing 129, 130, or 131 people with blood type O+ in 400 randomly selected people when $p = 0.35$.

Instead of standardizing 129 and 131, we will standardize 128.5 and 131.5:

(E)

$$Z_{left} = \frac{128.5 - 140}{9.5} = -1.263$$
$$Z_{right} = \frac{131.5 - 140}{9.5} = -0.895$$
$$P(-1.263 < Z < -0.895) = 0.0772$$

The probability 0.0772 is much closer to the true value of 0.0732 than the previous estimate of 0.0483 we calculated using normal approximation without the continuity correction.

It is always possible to apply the continuity correction when finding a normal approximation to the binomial distribution. However, when n is very large or when the interval is wide, the benefit of the modification is limited since the added area becomes negligible compared to the overall area being calculated.

Section summary

- $\binom{n}{x}$, the **binomial coefficient**, describes the number of combinations for arranging x successes among n trials. $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, where $n! = 1 \times 2 \times 3 \times ...n$, and 0!=0.
- The **binomial formula** can be used to find the probability that something happens *exactly x times in n trials*. Suppose the probability of a single trial being a success is p. Then the probability of observing exactly x successes in n independent trials is given by

$$\binom{n}{x} p^x (1-p)^{n-x} \quad = \quad \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

- To apply the binomial formula, the events must be **independent** from trial to trial. Additionally, n, the number of trials must be fixed in advance, and p, the probability of the event occurring in a given trial, must be the same for each trial.
- To use the binomial formula, first confirm that the binomial conditions are met. Next, identify the number of trials n, the number of times the event is to be a "success" x, and the probability that a single trial is a success p. Finally, plug these three numbers into the formula to get the probability of exactly x successes in n trials.
- To find a probability involving *at least* or *at most*, first determine if the scenario is binomial. If so, apply the binomial formula as many times as needed and add up the results. e.g. P(at least 3 Heads in 5 tosses of a fair coin) $=$ P(exactly 3 Heads) + P(exactly 4 Heads) + P(exactly 5 Heads), where each probability can be found using the binomial formula.
- The distribution of the *number of successes* in n independent trials gives rise to a **binomial distribution**. If X has a binomial distribution with parameters n and p, then $P(X = x) = \binom{n}{x} p^x (1-p)^n - x$, where $x = 0, 1, 2, 3 \ldots, n$.
- To write out a binomial probability **distribution table**, list all possible values for x, the number of successes, then use the binomial formula to find the probability of each of those values.
- If X follows a binomial distribution with parameters n and p, then:
  - The mean is given by $\mu_x = np$. (*center*)
  - The standard deviation is given by $\sigma_x = \sqrt{np(1-p)}$. (*spread*)
  - When $np \geq 10$ and $n(1-p) \geq 10$, the binomial distribution is approximately normal. (*shape*)

Exercises

3.51 Exploring combinations. A coin is tossed 5 times. How many sequences / combinations of Heads/Tails are there that have:

(a) Exactly 1 Tail?
(b) Exactly 4 Tails?
(c) Exactly 3 Tails?
(d) At least 3 Tails?

3.52 Political affiliation. Suppose that in a large population, 51% identify as Democrat. A researcher takes a random sample of 3 people.

(a) Use the binomial model to calculate the probability that two of them identify as Democrat.
(b) Write out all possible orderings of 3 people, 2 of whom identify as Democrat. Use these scenarios to calculate the same probability from part (a) but using the Addition Rule for disjoint events. Confirm that your answers from parts (a) and (b) match.
(c) If we wanted to calculate the probability that a random sample of 8 people will have 3 that identify as Democrat, briefly describe why the approach from part (b) would be more tedious than the approach from part (a).

3.53 Underage drinking, Part I. Data collected by the Substance Abuse and Mental Health Services Administration (SAMSHA) suggests that 69.7% of 18-20 year olds consumed alcoholic beverages in any given year.[86]

(a) Suppose a random sample of ten 18-20 year olds is taken. Is the use of the binomial distribution appropriate for calculating the probability that exactly six consumed alcoholic beverages? Explain.
(b) Calculate the probability that exactly 6 out of 10 randomly sampled 18- 20 year olds consumed an alcoholic drink.
(c) What is the probability that exactly four out of ten 18-20 year olds have *not* consumed an alcoholic beverage?
(d) What is the probability that at most 2 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?
(e) What is the probability that at least 1 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?

3.54 Chicken pox, Part I. The National Vaccine Information Center estimates that 90% of Americans have had chickenpox by the time they reach adulthood.[87]

(a) Suppose we take a random sample of 100 American adults. Is the use of the binomial distribution appropriate for calculating the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood? Explain.
(b) Calculate the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood.
(c) What is the probability that exactly 3 out of a new sample of 100 American adults have *not* had chickenpox in their childhood?
(d) What is the probability that at least 1 out of 10 randomly sampled American adults have had chickenpox?
(e) What is the probability that at most 3 out of 10 randomly sampled American adults have *not* had chickenpox?

3.55 Game of dreidel. A dreidel is a four-sided spinning top with the Hebrew letters *nun*, *gimel*, *hei*, and *shin*, one on each side. Each side is equally likely to come up in a single spin of the dreidel. Suppose you spin a dreidel three times. Calculate the probability of getting

(a) at least one *nun*?
(b) exactly 2 *nuns*?
(c) exactly 1 *hei*?
(d) at most 2 *gimels*?

Photo by Staccabees, cropped (http://flic.kr/p/7gLZTf) CC BY 2.0 license

[86] **webpage:alcohol**.
[87] **webpage:chickenpox**.

3.56 Sickle cell anemia. Sickle cell anemia is a genetic blood disorder where red blood cells lose their flexibility and assume an abnormal, rigid, "sickle" shape, which results in a risk of various complications. If both parents are carriers of the disease, then a child has a 25% chance of having the disease, 50% chance of being a carrier, and 25% chance of neither having the disease nor being a carrier. If two parents who are carriers of the disease have 3 children, what is the probability that

(a) two will have the disease?

(b) none will have the disease?

(c) at least one will neither have the disease nor be a carrier?

(d) the first child with the disease will the be 3^{rd} child?

3.57 Underage drinking, Part II.  We learned in Exercise 3.53 that about 70% of 18-20 year olds consumed alcoholic beverages in any given year. We now consider a random sample of fifty 18-20 year olds.

(a) How many people would you expect to have consumed alcoholic beverages? And with what standard deviation?

(b) Would you be surprised if there were 45 or more people who have consumed alcoholic beverages?

(c) What is the probability that 45 or more people in this sample have consumed alcoholic beverages? How does this probability relate to your answer to part (b)?

3.58 Chickenpox, Part II. We learned in Exercise 3.54 that about 90% of American adults had chickenpox before adulthood. We now consider a random sample of 120 American adults.

(a) How many people in this sample would you expect to have had chickenpox in their childhood? And with what standard deviation?

(b) Would you be surprised if there were 105 people who have had chickenpox in their childhood?

(c) What is the probability that 105 or fewer people in this sample have had chickenpox in their childhood? How does this probability relate to your answer to part (b)?

Chapter highlights

This chapter focused on understanding likelihood and chance variation, first by solving individual probability questions and then by investigating probability distributions.

The main probability techniques covered in this chapter are as follows:

- The **General Multiplication Rule** for **and** probabilities (intersection), along with the special case when events are **independent**.
- The **General Addition Rule** for **or** probabilities (union), along with the special case when events are **mutually exclusive**.
- The **Conditional Probability Rule**.
- Tree diagrams and **Bayes' Theorem** to solve more complex conditional problems.
- **Simulations** and the use of random digits to estimate probabilities.

Fundamental to all of these problems is understanding when events are independent and when they are mutually exclusive. Two events are **independent** when the outcome of one does not affect the outcome of the other, i.e. $P(A|B) = P(A)$. Two events are **mutually exclusive** when they cannot both happen together, i.e. $P(A \text{ and } B) = 0$.

Moving from solving individual probability questions to studying probability distributions helps us better understand chance processes and quantify expected chance variation.

- For a **discrete probability distribution**, the **sum** of the probabilities must equal 1.
- As with any distribution, one can calculate the mean and standard deviation of a probability distribution. In the context of a probability distribution, the **mean** and **standard deviation** describe the average and the typical deviation from the average, respectively, after many, many repetitions of the chance process.
- A probability distribution can be summarized by its **center** (mean, median), **spread** (SD, IQR), and **shape** (right skewed, left skewed, approximately symmetric).
- Adding a constant to every value in a probability distribution adds that value to the mean, but it does not affect the standard deviation. When multiplying every value by a constant, this multiplies the mean by the constant and it multiplies the standard deviation by the absolute value of the constant.
- The mean of the sum of two random variables equals the sum of the means. However, this is not true for standard deviations. Instead, when finding the standard deviation of a sum or difference of random variables, take the square root of the sum of each of the standard deviations squared.
- The **geometric distribution** provides a model for the number of trials until the first success, when the trials are independent.
- The **binomial distribution** provides a model for the number of successes in n independent trials.
- The geometric distribution is always right skewed. However, when the success-failure rule is met (at least 10 success and 10 failures), the binomial distribution can be modeled using a normal distribution with mean $= np$ and standard deviation $\sqrt{np(1-p)}$.

The study of probability is useful for measuring uncertainty and assessing risk. In addition, probability serves as the foundation for inference, providing a framework for evaluating when an outcome falls outside of the range of what would be expected by chance alone.

Chapter exercises

3.59 Grade distributions. Each row in the table below is a proposed grade distribution for a class. Identify each as a valid or invalid probability distribution, and explain your reasoning.

	Grades				
	A	B	C	D	F
(a)	0.3	0.3	0.3	0.2	0.1
(b)	0	0	1	0	0
(c)	0.3	0.3	0.3	0	0
(d)	0.3	0.5	0.2	0.1	-0.1
(e)	0.2	0.4	0.2	0.1	0.1
(f)	0	-0.1	1.1	0	0

3.60 Health coverage, frequencies. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table summarizes two variables for the respondents: health status and health coverage, which describes whether each respondent had health insurance.[88]

		Health Status					
		Excellent	Very good	Good	Fair	Poor	Total
Health Coverage	No	459	727	854	385	99	2,524
	Yes	4,198	6,245	4,821	1,634	578	17,476
	Total	4,657	6,972	5,675	2,019	677	20,000

(a) If we draw one individual at random, what is the probability that the respondent has excellent health and doesn't have health coverage?

(b) If we draw one individual at random, what is the probability that the respondent has excellent health or doesn't have health coverage?

3.61 HIV in Swaziland. Swaziland has the highest HIV prevalence in the world: 25.9% of this country's population is infected with HIV.[89] The ELISA test is one of the first and most accurate tests for HIV. For those who carry HIV, the ELISA test is 99.7% accurate. For those who do not carry HIV, the test is 92.6% accurate. If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

3.62 Twins. About 30% of human twins are identical, and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the probability that they are identical?

3.63 Cost of breakfast. Sally gets a cup of coffee and a muffin every day for breakfast from one of the many coffee shops in her neighborhood. She picks a coffee shop each morning at random and independently of previous days. The average price of a cup of coffee is \$1.40 with a standard deviation of 30¢ (\$0.30), the average price of a muffin is \$2.50 with a standard deviation of 15¢, and the two prices are independent of each other.

(a) What is the mean and standard deviation of the amount she spends on breakfast daily?

(b) What is the mean and standard deviation of the amount she spends on breakfast weekly (7 days)?

[88] **data:BRFSS2010**.
[89] **ciaFactBookHIV:2012**.

3.64 Scooping ice cream. Ice cream usually comes in 1.5 quart boxes (48 fluid ounces), and ice cream scoops hold about 2 ounces. However, there is some variability in the amount of ice cream in a box as well as the amount of ice cream scooped out. We represent the amount of ice cream in the box as X and the amount scooped out as Y. Suppose these random variables have the following means, standard deviations, and variances:

	mean	SD	variance
X	48	1	1
Y	2	0.25	0.0625

(a) An entire box of ice cream, plus 3 scoops from a second box is served at a party. How much ice cream do you expect to have been served at this party? What is the standard deviation of the amount of ice cream served?

(b) How much ice cream would you expect to be left in the box after scooping out one scoop of ice cream? That is, find the expected value of $X - Y$. What is the standard deviation of the amount left in the box?

(c) Using the context of this exercise, explain why we add variances when we subtract one random variable from another.

3.65 University admissions. Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the approximate probability that the university will not have enough dormitory room spots for the freshman class?

3.66 Speeding on the I-5, Part II. Exercise **??** states that the distribution of speeds of cars traveling on the Interstate 5 Freeway (I-5) in California is nearly normal with a mean of 72.6 miles/hour and a standard deviation of 4.78 miles/hour. The speed limit on this stretch of the I-5 is 70 miles/hour.

(a) A highway patrol officer is hidden on the side of the freeway. What is the probability that 5 cars pass and none are speeding? Assume that the speeds of the cars are independent of each other.

(b) On average, how many cars would the highway patrol officer expect to watch until the first car that is speeding? What is the standard deviation of the number of cars he would expect to watch?

Chapter 8

Introduction to linear regression

8.1 Line fitting, residuals, and correlation

8.2 Fitting a line by least squares regression

8.3 Transformations for skewed data

8.4 Inference for the slope of a regression line

Linear regression is a very powerful statistical technique. Many people have some familiarity with regression just from reading the news, where graphs with straight lines are overlaid on scatterplots. Linear models can be used to see trends and to make predictions.

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

8.1 Line fitting, residuals, and correlation

In this section, we investigate bivariate data. We examine criteria for identifying a linear model and introduce a new bivariate summary called *correlation*. We answer questions such as the following:

- How do we quantify the strength of the linear association between two numerical variables?
- What does it mean for two variables to have no association or to have a nonlinear association?
- Once we fit a model, how do we measure the error in the model's predictions?

Learning objectives

1. Distinguish between the data point y and the predicted value $\hat{y}$ based on a model.
2. Calculate a residual and draw a residual plot.
3. Interpret the standard deviation of the residuals.
4. Interpret the correlation coefficient and estimate it from a scatterplot.
5. Know and apply the properties of the correlation coefficient.

8.1.1 Fitting a line to data

Requests from twelve separate buyers were simultaneously placed with a trading company to purchase Target Corporation stock (ticker `TGT`, April 26th, 2012). We let x be the number of stocks to purchase and y be the total cost. Because the cost is computed using a linear formula, the linear fit is perfect, and the equation for the line is: $y = 5 + 57.49x$. If we know the number of stocks purchased, we can determine the cost based on this linear equation with no error. Additionally, we can say that each additional share of the stock cost \$57.49 and that there was a \$5 fee for the transaction.

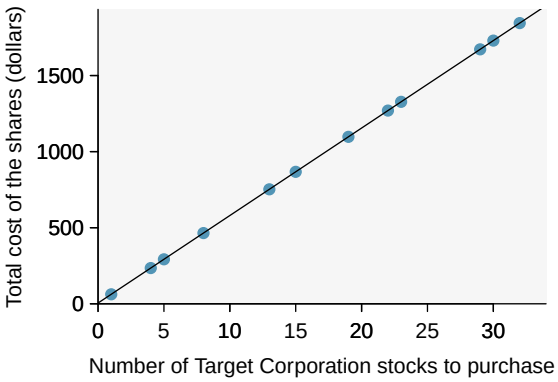


Figure 8.1: Total cost of a trade against number of shares purchased.

Perfect linear relationships are unrealistic in almost any natural process. For example, if we took family income (x), this value would provide some useful information about how much financial support a college may offer a prospective student (y). However, the prediction would be far from perfect, since other factors play a role in financial support beyond a family's income.

It is rare for all of the data to fall perfectly on a straight line. Instead, it's more common for data to appear as a *cloud of points*, such as those shown in Figure 8.2. In each case, the data fall around a straight line, even if none of the observations fall exactly on the line. The first plot shows a relatively strong downward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y. The second plot shows an upward trend that, while evident, is not as strong as the first. The last plot shows a very weak downward trend in the data, so slight we can hardly notice it.

In each of these examples, we can consider how to draw a "best fit line". For instance, we might wonder, should we move the line up or down a little, or should we tilt it more or less? As we move forward in this chapter, we will learn different criteria for line-fitting, and we will also learn about the uncertainty associated with estimates of model parameters.

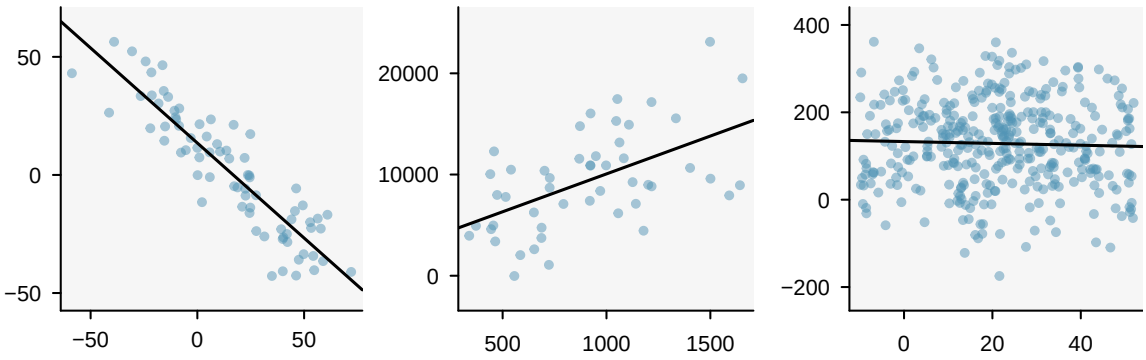


Figure 8.2: Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

We will also see examples in this chapter where fitting a straight line to the data, even if there is a clear relationship between the variables, is not helpful. One such case is shown in Figure 8.3 where there is a very strong relationship between the variables even though the trend is not linear.

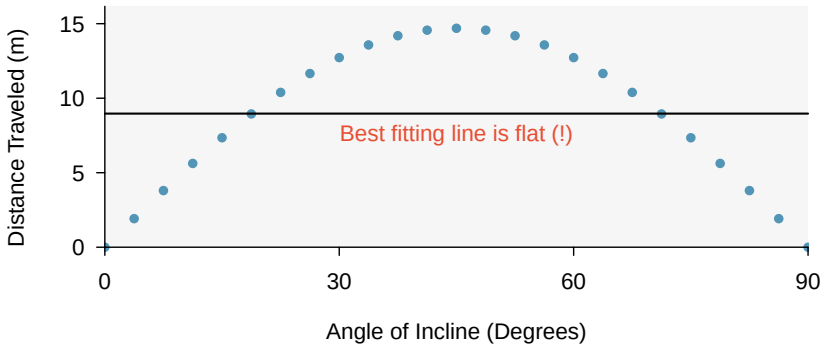


Figure 8.3: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

8.1.2 Using linear regression to predict possum head lengths

Brushtail possums are a marsupial that lives in Australia. A photo of one is shown in Figure 8.4. Researchers captured 104 of these animals and took body measurements before releasing the animals back into the wild. We consider two of these measurements: the total length of each possum, from head to tail, and the length of each possum's head.

Figure 8.5 shows a scatterplot for the head length and total length of the 104 possums. Each point represents a single point from the data.

Figure 8.4: The common brushtail possum of Australia.

Photo by Peter Firminger on Flickr: http://flic.kr/p/6aPTn CC BY 2.0 license.

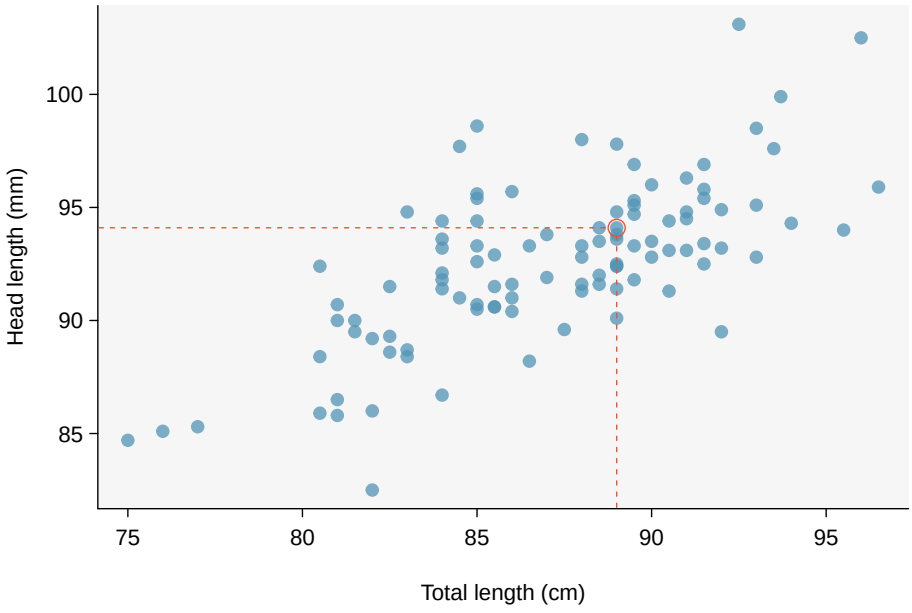


Figure 8.5: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1 mm and total length 89 cm is highlighted.

The head and total length variables are associated: possums with an above average total length also tend to have above average head lengths. While the relationship is not perfectly linear, it could be helpful to partially explain the connection between these variables with a straight line.

We want to describe the relationship between the head length and total length variables in the possum data set using a line. In this example, we will use the total length, x, to explain or predict a possum's head length, y. When we use x to predict y, we usually call x the **explanatory variable** or predictor variable, and we call y the **response variable**. We could fit the linear relationship by eye, as in Figure 8.6. The equation for this line is

$$\hat{y} = 41 + 0.59x$$

A "hat" on y is used to signify that this is a predicted value, not an observed value. We can use this line to discuss properties of possums. For instance, the equation predicts a possum with a total length of 80 cm will have a head length of

$$\begin{aligned} \hat{y} &= 41 + 0.59(80) \\ &= 88.2 \end{aligned}$$

The value $\hat{y}$ may be viewed as an average: the equation predicts that possums with a total length of 80 cm will have an average head length of 88.2 mm. The value $\hat{y}$ is also a prediction: absent further information about an 80 cm possum, this is our best prediction for a the head length of a single 80 cm possum.

8.1.3 Residuals

Residuals are the leftover variation in the response variable after fitting a model. Each observation will have a residual, and three of the residuals for the linear model we fit for the `possum` data are shown in Figure 8.6. If an observation is above the regression line, then its residual, the vertical distance from the observation to the line, is positive. Observations below the line have negative residuals. One goal in picking the right linear model is for these residuals to be as small as possible.

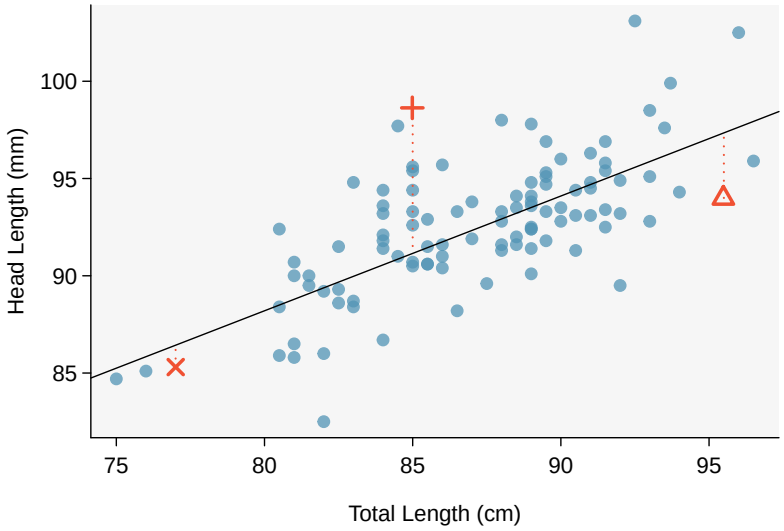


Figure 8.6: A reasonable linear model was fit to represent the relationship between head length and total length.

Let's look closer at the three residuals featured in Figure 8.6. The observation marked by an "×" has a small, negative residual of about -1; the observation marked by "+" has a large residual of about +7; and the observation marked by "△" has a moderate residual of about -4. The size of a residual is usually discussed in terms of its absolute value. For example, the residual for "△" is larger than that of "×" because $|-4|$ is larger than $|-1|$.

RESIDUAL: DIFFERENCE BETWEEN OBSERVED AND EXPECTED

The residual for a particular observation $(x,\ y)$ is the difference between the observed response and the response we would predict based on the model:

$$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y} \end{aligned}$$

We typically identify $\hat{y}$ by plugging x into the model.

EXAMPLE 8.1

The linear fit shown in Figure 8.6 is given as $\hat{y} = 41 + 0.59x$. Based on this line, compute and interpret the residual of the observation (77.0, 85.3). This observation is denoted by "×" on the plot. Recall that x is the total length measured in cm and y is head length measured in mm.

We first compute the predicted value based on the model:

$$\begin{aligned}\hat{y} &= 41 + 0.59x \\ &= 41 + 0.59(77.0) \\ &= 86.4\end{aligned}$$

(E)

Next we compute the difference of the actual head length and the predicted head length:

$$\begin{aligned}residual &= y - \hat{y} \\ &= 85.3 - 86.4 \\ &= -1.1\end{aligned}$$

The residual for this point is -1.1 mm, which is very close to the visual estimate of -1 mm. For this particular possum with total length of 77 cm, the model's prediction for its head length was 1.1 mm *too high*.

GUIDED PRACTICE 8.2

(G) If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?[1]

GUIDED PRACTICE 8.3

(G) Compute the residual for the observation $(95.5, 94.0)$, denoted by "△" in the figure, using the linear model: $\hat{y} = 41 + 0.59x$.[2]

Residuals are helpful in evaluating how well a linear model fits a data set. We often display the residuals in a **residual plot** such as the one shown in Figure 8.7. Here, the residuals are calculated for each x value, and plotted versus x. For instance, the point $(85.0, 98.6)$ had a residual of 7.45, so in the residual plot it is placed at $(85.0, 7.45)$. Creating a residual plot is sort of like tipping the scatterplot over so the regression line is horizontal.

From the residual plot, we can better estimate the **standard deviation of the residuals**, often denoted by the letter s. The standard deviation of the residuals tells us typical size of the residuals. As such, it is a measure of the typical deviation between the y values and the model predictions. In other words, it tells us the typical prediction error using the model.[3]

[1] If a model underestimates an observation, then the model estimate is below the actual. The residual, which is the actual observation value minus the model estimate, must then be positive. The opposite is true when the model overestimates the observation: the residual is negative.

[2] First compute the predicted value based on the model, then compute the residual.

$$\hat{y} = 41 + 0.59x = 41 + 0.59(95.50) = 97.3$$

$$residual = y - \hat{y} = 94.0 - 97.3 = -3.3$$

The residual is -3.3, so the model *overpredicted* the head length for this possum by 3.3 mm.

[3] The standard deviation of the residuals is calculated as: $s = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$.

EXAMPLE 8.4

Estimate the standard deviation of the residuals for predicting head length from total length using the line: $\hat{y} = 41 + 0.59x$ using Figure 8.7. Also, interpret the quantity in context.

(E)

To estimate this graphically, we use the residual plot. The approximate 68, 95 rule for standard deviations applies. Approximately 2/3 of the points are within ± 2.5 and approximately 95% of the points are within ± 5, so 2.5 is a good estimate for the standard deviation of the residuals. The typical error when predicting head length using this model is about 2.5 mm.

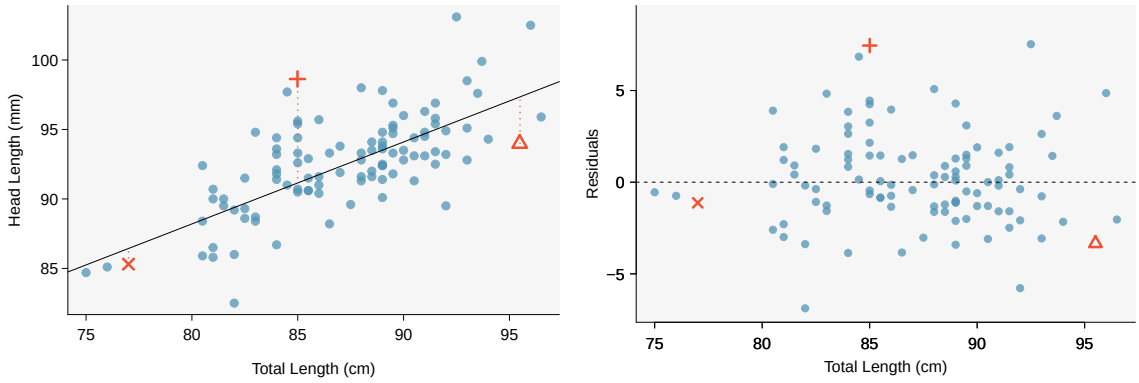


Figure 8.7: Left: Scatterplot of head length versus total length for 104 brushtail possums. Three particular points have been highlighted. Right: Residual plot for the model shown in left panel.

STANDARD DEVIATION OF THE RESIDUALS

The standard deviation of the residuals, often denoted by the letter s, tells us the typical error in the predictions using the regression model. It can be estimated from a residual plot.

EXAMPLE 8.5

One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model. Figure 8.8 shows three scatterplots with linear models in the first row and residual plots in the second row. Can you identify any patterns remaining in the residuals?

In the first data set (first column), the residuals show no obvious patterns. The residuals appear to be scattered randomly around the dashed line that represents 0.

(E) The second data set shows a pattern in the residuals. There is some curvature in the scatterplot, which is more obvious in the residual plot. We should not use a straight line to model these data. Instead, a more advanced technique should be used.

The last plot shows very little upwards trend, and the residuals also show no obvious patterns. It is reasonable to try to fit a linear model to the data. However, it is unclear whether there is statistically significant evidence that the slope parameter is different from zero. The slope of the sample regression line is not zero, but we might wonder if this could be due to random variation. We will address this sort of scenario in Section 8.4.

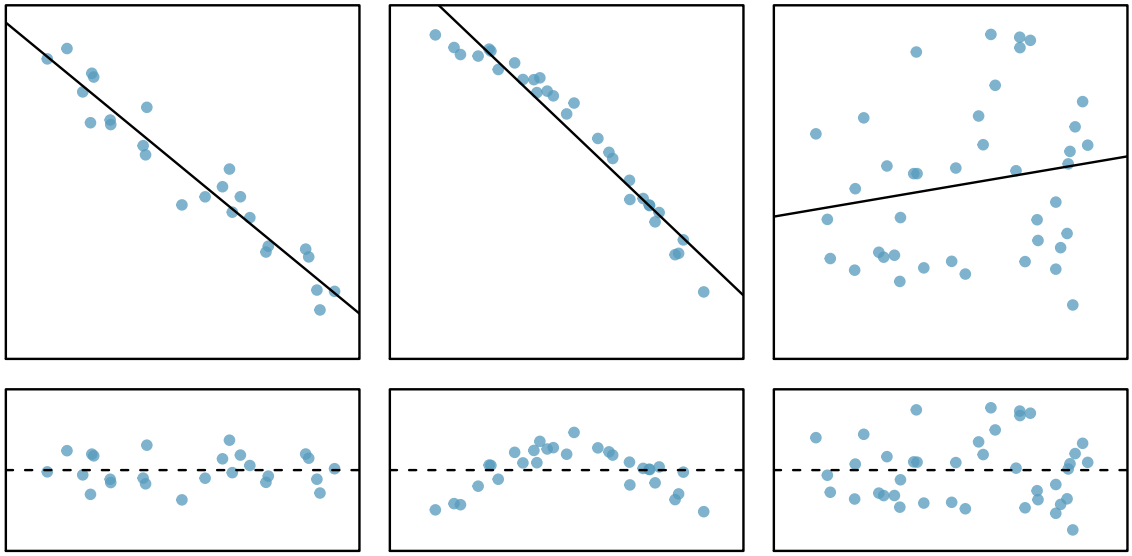


Figure 8.8: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

8.1.4 Describing linear relationships with correlation

When a linear relationship exists between two variables, we can quantify the strength and direction of the linear relation with the correlation coefficient, or just **correlation** for short. Figure 8.9 shows eight plots and their corresponding correlations.

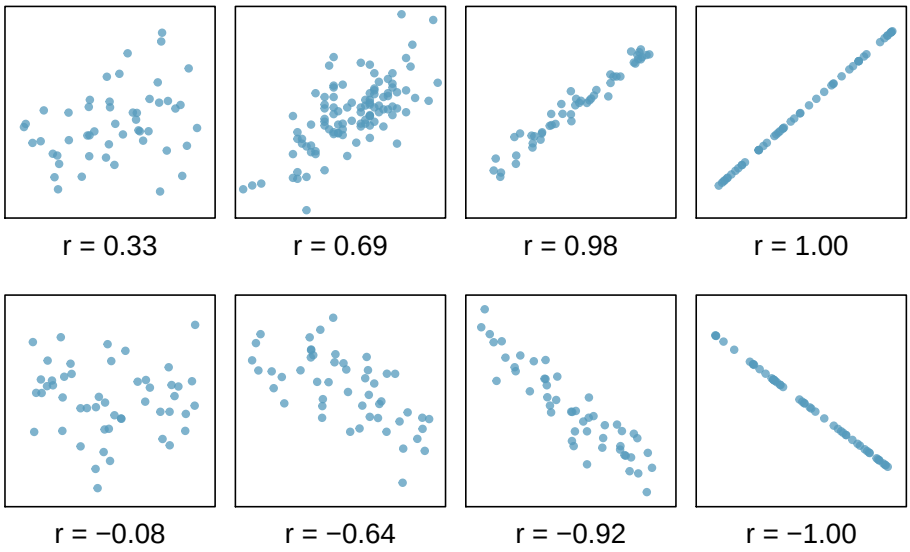


Figure 8.9: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

Only when the relationship is perfectly linear is the correlation either -1 or 1. If the linear relationship is strong and positive, the correlation will be near $+1$. If it is strong and negative, it will be near -1. If there is no apparent linear relationship between the variables, then the correlation will be near zero.

CORRELATION MEASURES THE STRENGTH OF A LINEAR RELATIONSHIP

Correlation, which always takes values between -1 and 1, describes the direction and strength of the linear relationship between two numerical variables. The strength can be strong, moderate, or weak.

We compute the correlation using a formula, just as we did with the sample mean and standard deviation. Formally, we can compute the correlation for observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) using the formula

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable. This formula is rather complex, and we generally perform the calculations on a computer or calculator. We can note, though, that the computation involves taking, for each point, the product of the Z-scores that correspond to the x and y values.

EXAMPLE 8.6

Take a look at Figure 8.6 on page 174. How would the correlation between head length and total body length of possums change if head length were measured in cm rather than mm? What if head length were measured in inches rather than mm?

(E) Here, changing the units of y corresponds to multiplying all the y values by a certain number. This would change the mean and the standard deviation of y, but it would not change the correlation. To see this, imagine dividing every number on the vertical axis by 10. The units of y are now in cm rather than in mm, but the graph has remain exactly the same. The units of y have changed, by the relative distance of the y values about the mean are the same; that is, the Z-scores corresponding to the y values have remained the same.

CHANGING UNITS OF X AND Y DOES NOT AFFECT THE CORRELATION

The correlation, r, between two variables is not dependent upon the units in which the variables are recorded. Correlation itself has no units.

Correlation is intended to quantify the strength of a linear trend. Nonlinear trends, even when strong, sometimes produce correlations that do not reflect the strength of the relationship; see three such examples in Figure 8.10.

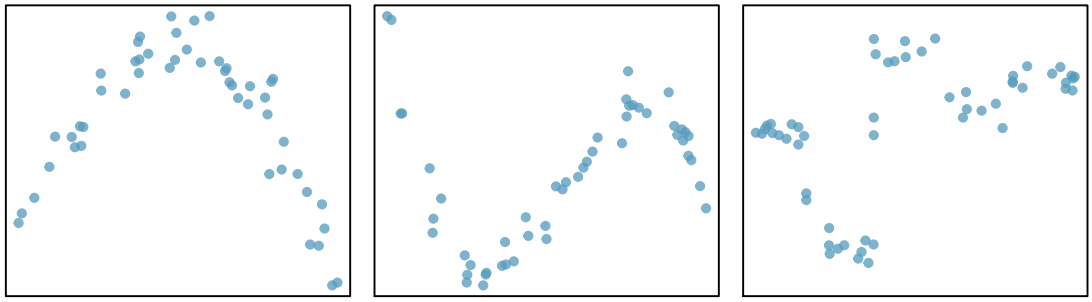


Figure 8.10: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, the correlation is not very strong, and the relationship is not linear.

GUIDED PRACTICE 8.7

(G) It appears no straight line would fit any of the datasets represented in Figure 8.10. Try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in your fit.[4]

[4]We'll leave it to you to draw the lines. In general, the lines you draw should be close to most points and reflect overall trends in the data.

EXAMPLE 8.8

Consider the four scatterplots in Figure 8.11. In which scatterplot is the correlation between x and y the strongest?

(E) All four data sets have the exact same correlation of $r = 0.816$ as well as the same equation for the best fit line! This group of four graphs, known as Anscombe's Quartet, remind us that knowing the value of the correlation does not tell us what the corresponding scatterplot looks like. It is always important to first graph the data. Investigate Anscombe's Quartet in Desmos: https://www.desmos.com/calculator/paknt6oneh.

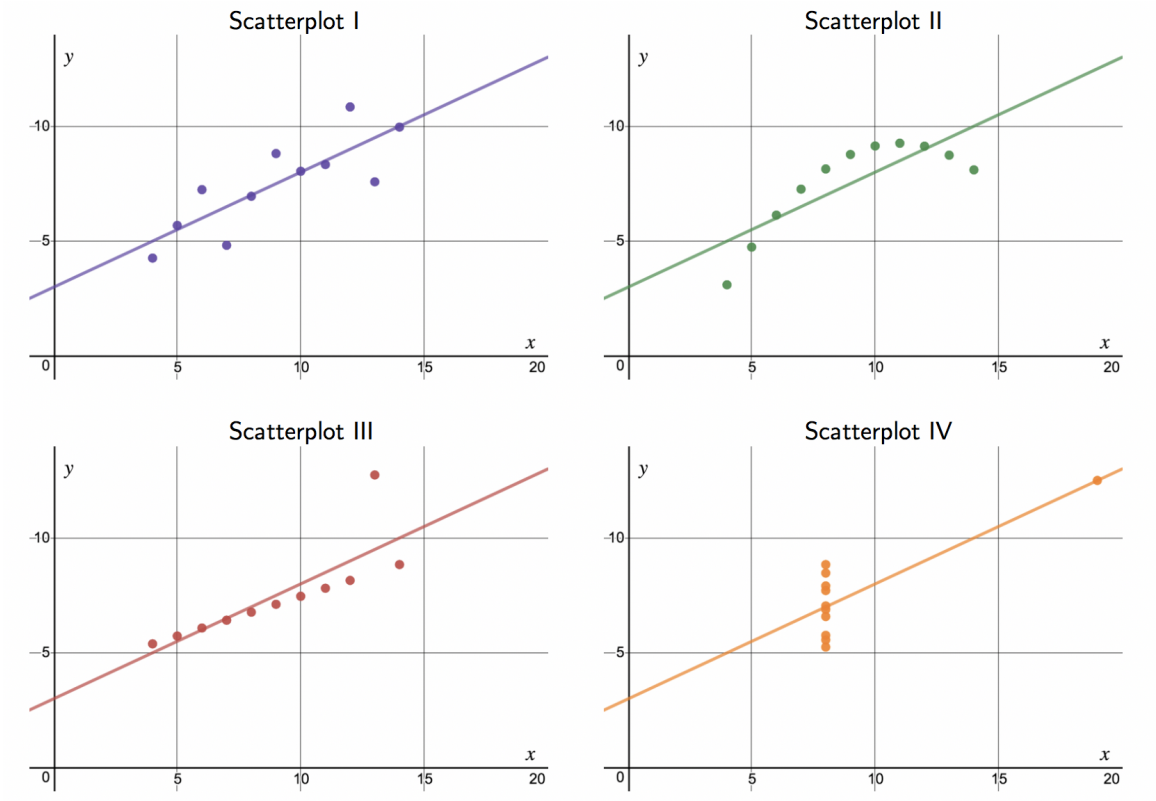


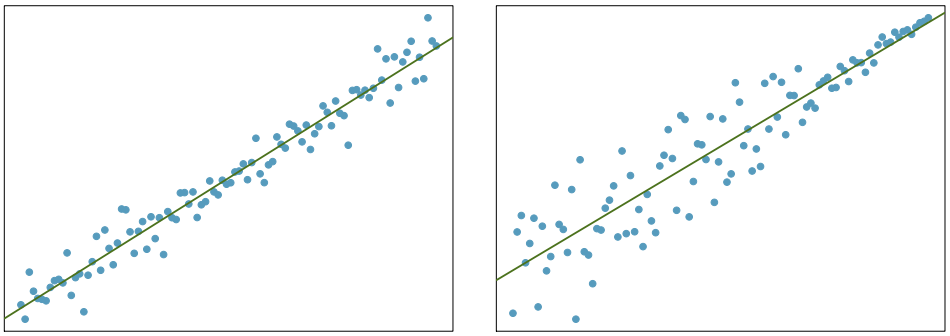
Figure 8.11: Four scatterplots from Desmos with best fit line drawn in.

Section summary

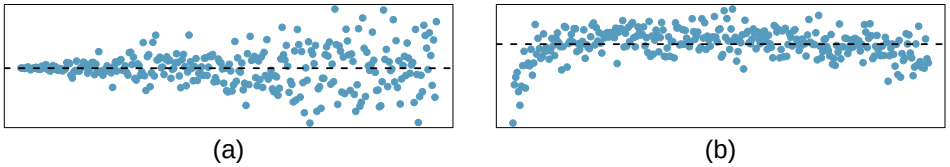
- In Chapter 2 we introduced a bivariate display called a **scatterplot**, which shows the relationship between two numerical variables. When we use x to predict y, we call x the **explanatory variable** or predictor variable, and we call y the **response variable**.
- A linear model for bivariate numerical data can be useful for prediction when the association between the variables follows a constant, linear trend. Linear models should not be used if the trend between the variables is curved.
- When we write a linear model, we use $\hat{y}$ to indicate that it is the model or the prediction. The value $\hat{y}$ can be understood as a **prediction** for y based on a given x, or as an **average** of the y values for a given x.
- The **residual** is the **error** between the true value and the modeled value, computed as $y - \hat{y}$. The order of the difference matters, and the sign of the residual will tell us if the model overpredicted or underpredicted a particular data point.
- The symbol s in a linear model is used to denote the standard deviation of the residuals, and it measures the typical prediction error by the model.
- A **residual plot** is a scatterplot with the residuals on the vertical axis. The residuals are often plotted against x on the horizontal axis, but they can also be plotted against y, $\hat{y}$, or other variables. Two important uses of a residual plot are the following.
  - Residual plots help us see patterns in the data that may not have been apparent in the scatterplot.
  - The standard deviation of the residuals is easier to estimate from a residual plot than from the original scatterplot.
- **Correlation**, denoted with the letter r, measures the strength and direction of a linear relationship. The following are some important facts about correlation.
  - The value of r is always between -1 and 1, inclusive, with an $r = -1$ indicating a perfect negative relationship (points fall exactly along a line that has negative slope) and an $r = 1$ indicating a perfect positive relationship (points fall exactly along a line that has positive slope).
  - An $r = 0$ indicates no *linear* association between the variables, though there may well exist a quadratic or other type of association.
  - Just like Z-scores, the correlation has no units. Changing the units in which x or y are measured does not affect the correlation.
  - Correlation is sensitive to outliers. Adding or removing a single point can have a big effect on the correlation.
  - As we learned previously, correlation is not causation. Even a very strong correlation cannot prove causation; only a well-designed, controlled, randomized experiment can prove causation.

Exercises

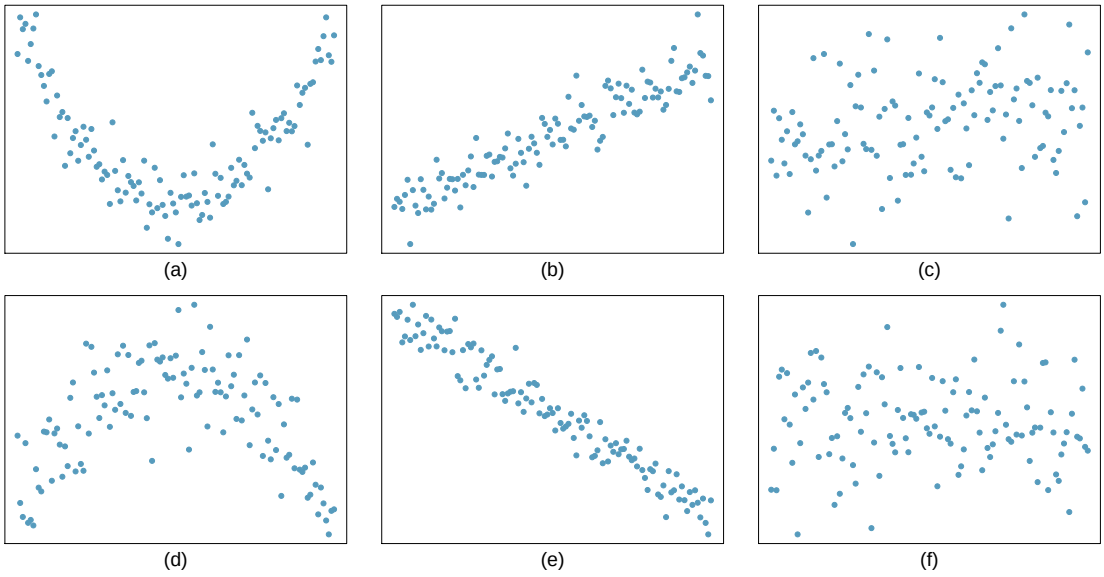
8.1 Visualize the residuals. The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.



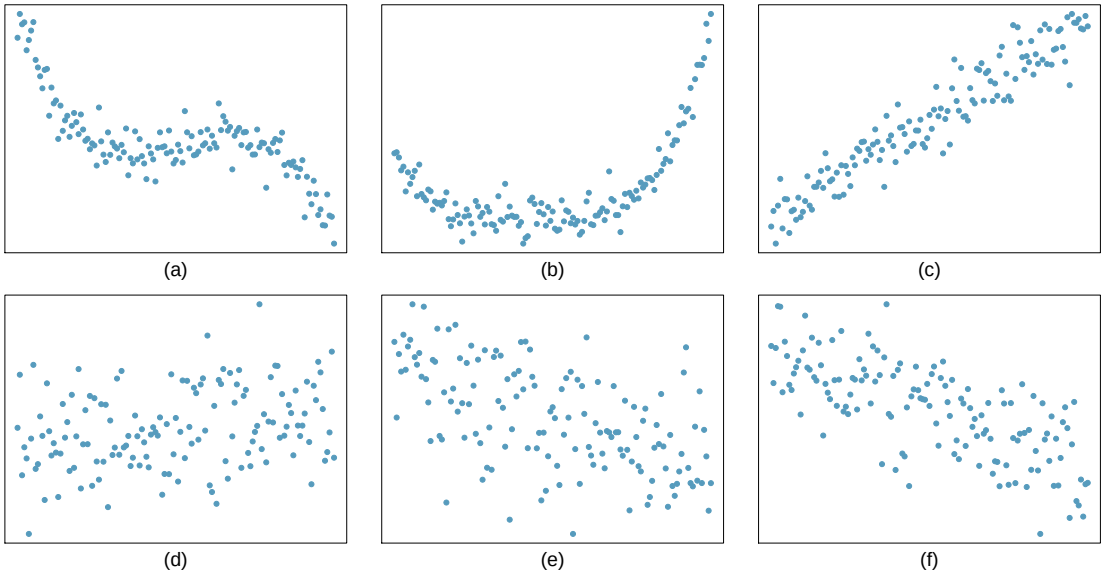
8.2 Trends in the residuals. Shown below are two plots of residuals remaining after fitting a linear model to two different sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.



8.3 Identify relationships, Part I. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.



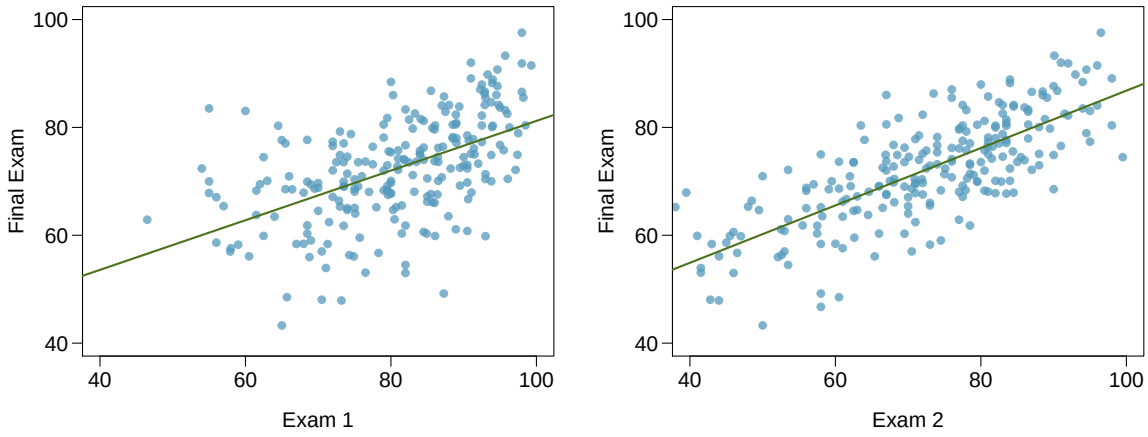
8.4 Identify relationships, Part II. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.



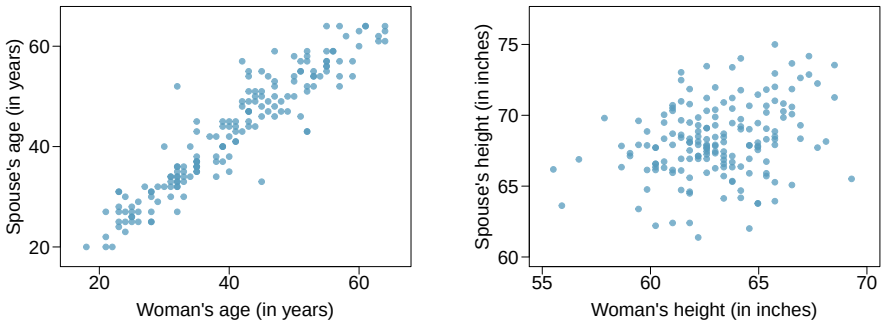
8.5 Exams and grades. The two scatterplots below show the relationship between final and mid-semester exam grades recorded during several years for a Statistics course at a university.

(a) Based on these graphs, which of the two exams has the strongest correlation with the final exam grade? Explain.

(b) Can you think of a reason why the correlation between the exam you chose in part (a) and the final exam is higher?



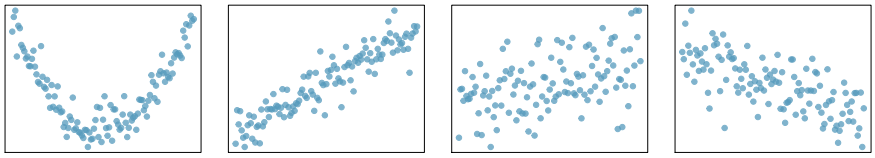
8.6 Spouses, Part I. The Great Britain Office of Population Census and Surveys once collected data on a random sample of 170 married women in Britain, recording the age (in years) and heights (converted here to inches) of the women and their spouses.[5] The scatterplot on the left shows the spouse's age plotted against the woman's age, and the plot on the right shows spouse's height plotted against the woman's height.



(a) Describe the relationship between the ages of women in the sample and their spouses' ages.
(b) Describe the relationship between the heights of women in the sample and their spouses' heights.
(c) Which plot shows a stronger correlation? Explain your reasoning.
(d) Data on heights were originally collected in centimeters, and then converted to inches. Does this conversion affect the correlation between heights of women in the sample and their spouses' heights?

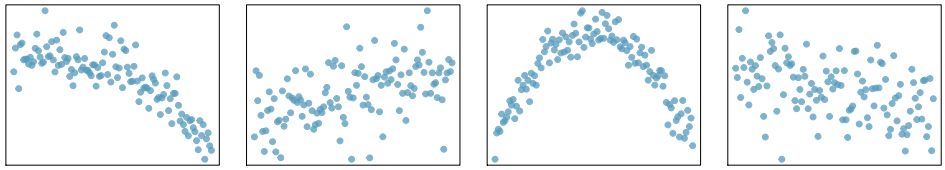
8.7 Match the correlation, Part I. Match each correlation to the corresponding scatterplot.

(a) $r = -0.7$
(b) $r = 0.45$
(c) $r = 0.06$
(d) $r = 0.92$

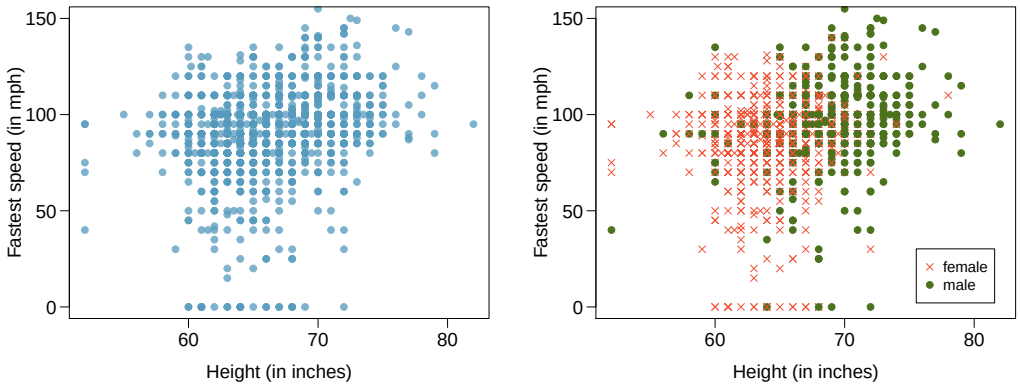


8.8 Match the correlation, Part II. Match each correlation to the corresponding scatterplot.

(a) $r = 0.49$
(b) $r = -0.48$
(c) $r = -0.03$
(d) $r = -0.85$



8.9 Speed and height. 1,302 UCLA students were asked to fill out a survey where they were asked about their height, fastest speed they have ever driven, and gender. The scatterplot on the left displays the relationship between height and fastest speed, and the scatterplot on the right displays the breakdown by gender in this relationship.



(a) Describe the relationship between height and fastest speed.
(b) Why do you think these variables are positively associated?
(c) What role does gender play in the relationship between height and fastest driving speed?

[5] **Hand:1994**.

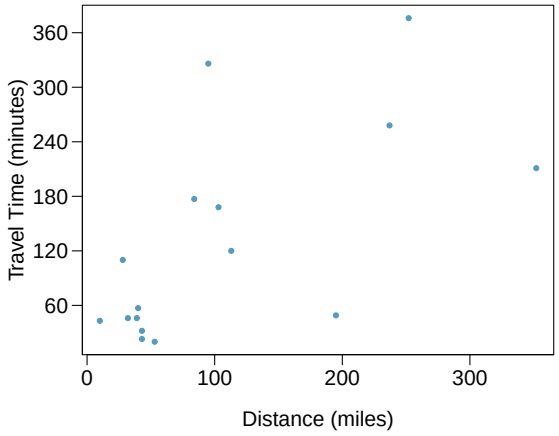
8.10 Guess the correlation. Eduardo and Rosie are both collecting data on number of rainy days in a year and the total rainfall for the year. Eduardo records rainfall in inches and Rosie in centimeters. How will their correlation coefficients compare?

8.11 The Coast Starlight, Part I. The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).

(a) Describe the relationship between distance and travel time.

(b) How would the relationship change if travel time was instead measured in hours, and distance was instead measured in kilometers?

(c) Correlation between travel time (in miles) and distance (in minutes) is $r = 0.636$. What is the correlation between travel time (in kilometers) and distance (in hours)?

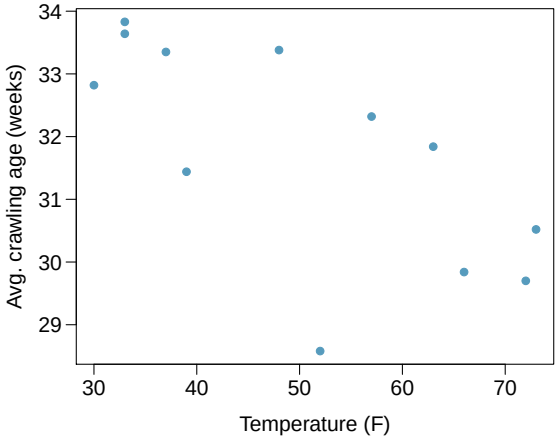


8.12 Crawling babies, Part I. A study conducted at the University of Denver investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months.[6] Infants born during the study year were split into twelve groups, one for each birth month. We consider the average crawling age of babies in each group against the average temperature when the babies are six months old (that's when babies often begin trying to crawl). Temperature is measured in degrees Fahrenheit (°F) and age is measured in weeks.

(a) Describe the relationship between temperature and crawling age.

(b) How would the relationship change if temperature was measured in degrees Celsius (°C) and age was measured in months?

(c) The correlation between temperature in °F and age in weeks was $r = -0.70$. If we converted the temperature to °C and age to months, what would the correlation be?

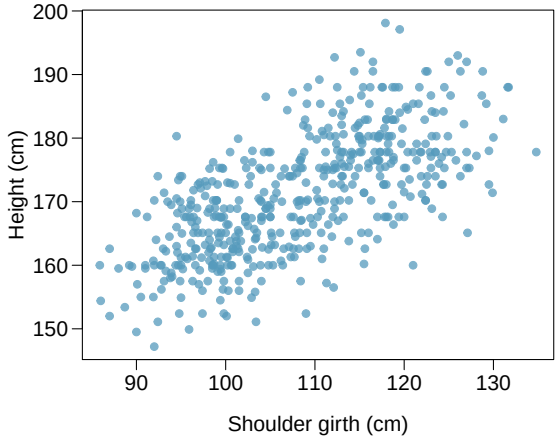


[6] **Benson:1993**.

8.13 Body measurements, Part I. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals.[7] The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.

(a) Describe the relationship between shoulder girth and height.

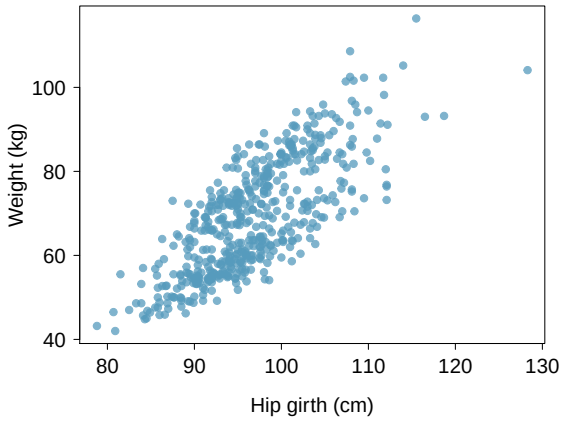
(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?



8.14 Body measurements, Part II. The scatterplot below shows the relationship between weight measured in kilograms and hip girth measured in centimeters from the data described in Exercise 8.13.

(a) Describe the relationship between hip girth and weight.

(b) How would the relationship change if weight was measured in pounds while the units for hip girth remained in centimeters?



8.15 Correlation, Part I. What would be the correlation between the ages of a set of women and their spouses if the set of women always married someone who was

(a) 3 years younger than themselves?

(b) 2 years older than themselves?

(c) half as old as themselves?

8.16 Correlation, Part II. What would be the correlation between the annual salaries of males and females at a company if for a certain type of position men always made

(a) $5,000 more than women?

(b) 25% more than women?

(c) 15% less than women?

[7] **Heinz:2003**.

8.2 Fitting a line by least squares regression

In this section, we answer the following questions:

- How well can we predict financial aid based on family income for a particular college?
- How does one find, interpret, and apply the least squares regression line?
- How do we measure the fit of a model and compare different models to each other?
- Why do models sometimes make predictions that are ridiculous or impossible?

Learning objectives

1. Calculate the slope and y-intercept of the least squares regression line using the relevant summary statistics. Interpret these quantities in context.
2. Understand why the least squares regression line is called the least squares regression line.
3. Interpret the explained variance R^2.
4. Understand the concept of extrapolation and why it is dangerous.
5. Identify outliers and influential points in a scatterplot.

8.2.1 An objective measure for finding the best line

Fitting linear models by eye is open to criticism since it is based on an individual preference. In this section, we use *least squares regression* as a more rigorous approach.

This section considers family income and gift aid data from a random sample of fifty students in the freshman class of Elmhurst College in Illinois.[8] Gift aid is financial aid that does not need to be paid back, as opposed to a loan. A scatterplot of the data is shown in Figure 8.12 along with two linear fits. The lines follow a negative trend in the data; students who have higher family incomes tended to have lower gift aid from the university.

We begin by thinking about what we mean by "best". Mathematically, we want a line that has small residuals. Perhaps our criterion could minimize the sum of the residual magnitudes:

$$|y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + \cdots + |y_n - \hat{y}_n|$$

which we could accomplish with a computer program. The resulting dashed line shown in Figure 8.12 demonstrates this fit can be quite reasonable. However, a more common practice is to choose the line that minimizes the sum of the squared residuals:

$$(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2$$

The line that minimizes the sum of the squared residuals is represented as the solid line in Figure 8.12. This is commonly called the **least squares line**.

Both lines seem reasonable, so why do data scientists prefer the least squares regression line? One reason is that it is easier to compute by hand and in most statistical software. Another, and more compelling, reason is that in many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by 2. Squaring the residuals accounts for this discrepancy.

[8] These data were sampled from a table of data for all freshmen from the 2011 class at Elmhurst College that accompanied an article titled *What Students Really Pay to Go to College* published online by *The Chronicle of Higher Education*: chronicle.com/article/What-Students-Really-Pay-to-Go/131435

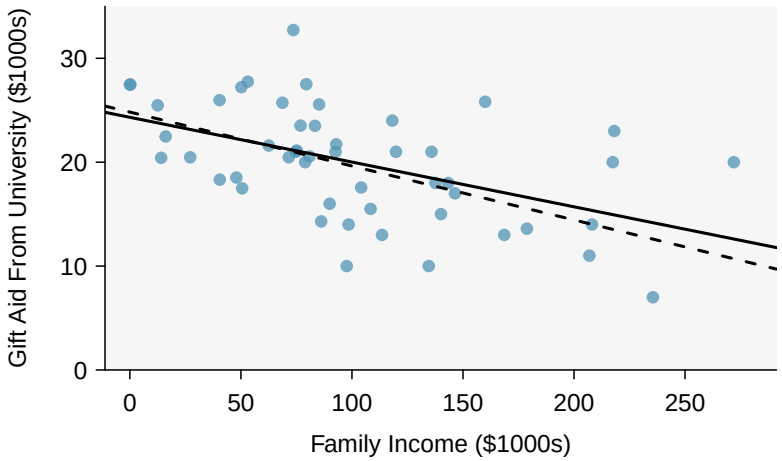


Figure 8.12: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are fit to the data, the solid line being the *least squares line*.

In Figure 8.13, we imagine the squared error about a line as actual squares. The least squares regression line minimizes the sum of the *areas* of these squared errors. In the figure, the sum of the squared error is $4 + 1 + 1 = 6$. There is no other line about which the sum of the squared error will be smaller.

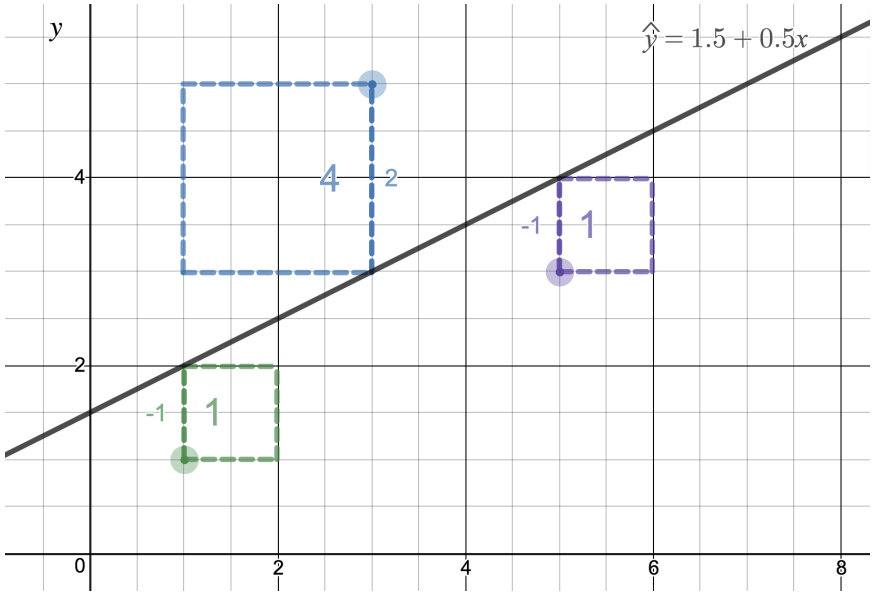


Figure 8.13: A visualization of least squares regression using Desmos. Try out this and other interactive Desmos activities at openintro.org/ahss/desmos.

8.2.2 Finding the least squares line

For the Elmhurst College data, we could fit a least squares regression line for predicting gift aid based on a student's family income and write the equation as:

$$\widehat{aid} = a + b \times family\_income$$

Here a is the y-intercept of the least squares regression line and b is the slope of the least squares regression line. a and b are both statistics that can be calculated from the data. In the next section we will consider the corresponding parameters that they statistics attempt to estimate.

We can enter all of the data into a statistical software package and easily find the values of a and b. However, we can also calculate these values by hand, using only the summary statistics.

- The slope of the least squares line is given by

$$b = r\frac{s_y}{s_x}$$

  where r is the correlation between the variables x and y, and s_x and s_y are the sample standard deviations of x, the explanatory variable, and y, the response variable.

- The point of averages $(\bar{x}, \bar{y})$ is always on the least squares line. Plugging this point in for x and y in the least squares equation and solving for a gives

$$\bar{y} = a + b\bar{x} \qquad\qquad a = \bar{y} - b\bar{x}$$

FINDING THE SLOPE AND INTERCEPT OF THE LEAST SQUARES REGRESSION LINE

The least squares regression line for predicting y based on x can be written as: $\hat{y} = a + bx$.

$$b = r\frac{s_y}{s_x} \qquad \bar{y} = a + b\bar{x}$$

We first find b, the slope, and then we solve for a, the y-intercept.

GUIDED PRACTICE 8.9

G Figure 8.14 shows the sample means for the family income and gift aid as \$101,800 and \$19,940, respectively. Plot the point $(101.8, 19.94)$ on Figure 8.12 to verify it falls on the least squares line (the solid line).[9]

	family income, in \$1000s ("x")	gift aid, in \$1000s ("y")
mean	$\bar{x} = 101.8$	$\bar{y} = 19.94$
sd	$s_x = 63.2$	$s_y = 5.46$
		$r = -0.499$

Figure 8.14: Summary statistics for family income and gift aid.

[9] If you need help finding this location, draw a straight line up from the x-value of 100 (or thereabout). Then draw a horizontal line at 20 (or thereabout). These lines should intersect on the least squares line.

EXAMPLE 8.10

Using the summary statistics in Figure 8.14, find the equation of the least squares regression line for predicting gift aid based on family income.

E

$$b = r\frac{s_y}{s_x} = (-0.499)\frac{5.46}{63.2} = -0.0431$$
$$a = \bar{y} - b\bar{x} = 19.94 - (-0.0431)(101.8) = 24.3$$

$$\hat{y} = 24.3 - 0.0431x \qquad \text{or} \qquad \widehat{aid} = 24.3 - 0.0431 \times family\_income$$

EXAMPLE 8.11

Say we wanted to predict a student's family income based on the amount of gift aid that they received. Would this least squares regression line be the following?

$$aid = 24.3 - 0.0431 \times \widehat{family\_income}$$

E

No. The equation we found was for predicting aid, not for predicting family income. We would have to calculate a new regression line, letting y be *family_income* and x be *aid*. This would give us:

$$b = r\frac{s_y}{s_x} = (-0.499)\frac{63.2}{5.46} = -5.776$$
$$a = \bar{y} - b\bar{x} = 19.94 - (-5.776)(101.8) = 607.9$$

$$\hat{y} = 607.3 - 5.776x \qquad \text{or} \qquad \widehat{family\_income} = 607.3 - 5.776 \times aid$$

We mentioned earlier that a computer is usually used to compute the least squares line. A summary table based on computer output is shown in Figure 8.15 for the Elmhurst College data. The first column of numbers provides estimates for b_0 and b_1, respectively. Compare these to the result from Example 8.2.2.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.3193	1.2915	18.83	0.0000
family_income	-0.0431	0.0108	-3.98	0.0002

Figure 8.15: Summary of least squares fit for the Elmhurst College data. Compare the parameter estimates in the first column to the results of Guided Practice 8.2.2.

EXAMPLE 8.12

Examine the second, third, and fourth columns in Figure 8.15. Can you guess what they represent?

E

We'll look at the second row, which corresponds to the slope. The first column, Estimate = -0.0431, tells us our best estimate for the slope of the population regression line. We call this point estimate b. The second column, Std. Error = 0.0108, is the standard error of this point estimate. The third column, t value = -3.98, is the T test statistic for the null hypothesis that the slope of the population regression line = 0. The last column, Pr(>|t|) = 0.0002, is the p-value for this two-sided T-test. We will get into more of these details in Section 8.4.

EXAMPLE 8.13

Suppose a high school senior is considering Elmhurst College. Can she simply use the linear equation that we have found to calculate her financial aid from the university?

(E) No. Using the equation will provide a prediction or estimate. However, as we see in the scatterplot, there is a lot of variability around the line. While the linear equation is good at capturing the trend in the data, there will be significant error in predicting an individual student's aid. Additionally, the data all come from one freshman class, and the way aid is determined by the university may change from year to year.

8.2.3 Interpreting the coefficients of a regression line

Interpreting the coefficients in a regression model is often one of the most important steps in the analysis.

EXAMPLE 8.14

The slope for the Elmhurst College data for predicting gift aid based on family income was calculated as -0.0431. Intepret this quantity in the context of the problem.

(E) You might recall from an algebra course that slope is change in y over change in x. Here, both x and y are in thousands of dollars. So if x is one unit or one thousand dollars higher, the line will predict that y will change by 0.0431 thousand dollars. In other words, for each additional thousand dollars of family income, *on average*, students receive 0.0431 thousand, or \$43.10 *less* in gift aid. Note that a higher family income corresponds to less aid because the slope is negative.

EXAMPLE 8.15

The y-intercept for the Elmhurst College data for predicting gift aid based on family income was calculated as 24.3. Intepret this quantity in the context of the problem.

(E) The intercept a describes the predicted value of y when $x = 0$. The *predicted* gift aid is 24.3 thousand dollars if a student's family has no income. The meaning of the intercept is relevant to this application since the family income for some students at Elmhurst is \$0. In other applications, the intercept may have little or no practical value if there are no observations where x is near zero. Here, it would be acceptable to say that the *average* gift aid is 24.3 thousand dollars among students whose family have 0 dollars in income.

INTERPRETING COEFFICIENTS IN A LINEAR MODEL

- The slope, b, describes the *average* increase or decrease in the y variable if the explanatory variable x is one unit larger.
- The y-intercept, a, describes the predicted outcome of y if $x = 0$. The linear model must be valid all the way to $x = 0$ for this to make sense, which in many applications is not the case.

GUIDED PRACTICE 8.16

In the previous chapter, we encountered a data set that compared the price of new textbooks for UCLA courses at the UCLA Bookstore and on Amazon. We fit a linear model for predicting price at UCLA Bookstore from price on Amazon and we get:

$$\hat{y} = 1.86 + 1.03x$$

where x is the price on Amazon and y is the price at the UCLA bookstore. Interpret the coefficients in this model and discuss whether the interpretations make sense in this context.[10]

GUIDED PRACTICE 8.17

Can we conclude that if Amazon raises the price of a textbook by 1 dollar, the UCLA Bookstore will raise the price of the textbook by \$1.03?[11]

EXERCISE CAUTION WHEN INTERPRETING COEFFICIENTS OF A LINEAR MODEL

- The slope tells us only the *average* change in y for each unit change in x; it does not tell us how much y might change based on a change in x for any particular *individual*. Moreover, in most cases, the slope cannot be interpreted in a causal way.
- When a value of $x = 0$ doesn't make sense in an application, then the interpretation of the y-intercept won't have any practical meaning.

8.2.4 Extrapolation is treacherous

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6^{th} it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on.

Stephen Colbert
April 6th, 2010 [12]

Linear models can be used to approximate the relationship between two variables. However, these models have real limitations. Linear regression is simply a modeling framework. The truth is almost always much more complex than our simple line. For example, we do not know how the data outside of our limited window will behave.

[10] The y-intercept is 1.86 and the units of y are in dollars. This tells us that when a textbook costs 0 dollars on Amazon, the predicted price of the textbook at the UCLA Bookstore is 1.86 dollars. This does not make sense as Amazon does not sell any \$0 textbooks. The slope is 1.03, with units (dollars)/(dollars). On average, for every extra dollar that a book costs on Amazon, it costs an extra 1.03 dollars at the UCLA Bookstore. This interpretation does make sense in this context.

[11] No. The slope describes the overall trend. This is observational data; a causal conclusion cannot be drawn. Remember, a causal relationship can only be concluded by a well-designed randomized, controlled experiment. Additionally, there may be large variation in the points about the line. The slope does not tell us how much y might change based on a change in x for a particular textbook.

[12] www.cc.com/video-clips/l4nkoq/

EXAMPLE 8.18

Use the model $\widehat{aid} = 24.3 - 0.0431 \times family\_income$ to estimate the aid of another freshman student whose family had income of \$1 million.

Recall that the units of family income are in \$1000s, so we want to calculate the aid for $family\_income = 1000$:

E

$$\widehat{aid} = 24.3 - 0.0431 \times family\_income$$
$$\widehat{aid} = 24.3 - 0.431(1000) = -18.8$$

The model predicts this student will have -\$18,800 in aid (!). Elmhurst College cannot (or at least does not) require any students to pay extra on top of tuition to attend.

Using a model to predict y-values for x-values outside the domain of the original data is called **extrapolation**. Generally, a linear model is only an approximation of the real relationship between two variables. If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analyzed.

8.2.5 Using R^2 to describe the strength of a fit

We evaluated the strength of the linear relationship between two variables earlier using the correlation, r. However, it is more common to explain the fit of a model using R^2, called **R-squared** or the **explained variance**. If provided with a linear model, we might like to describe how closely the data cluster around the linear fit.

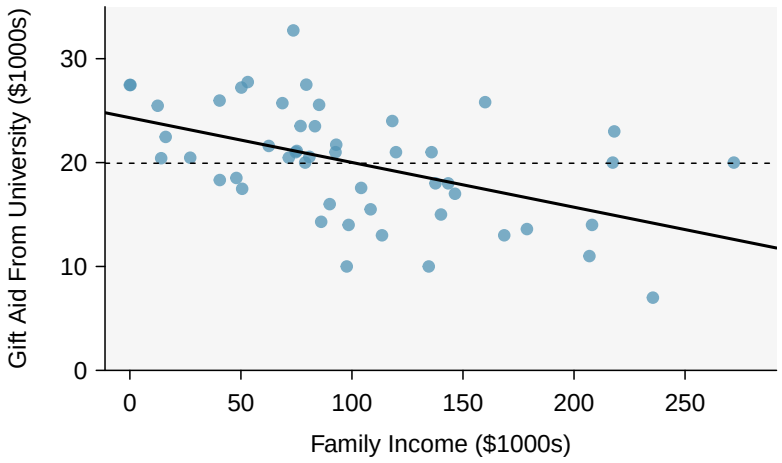


Figure 8.16: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College, shown with the least squares regression line ($\hat{y}$) and the average line ($\bar{y}$).

We are interested in how well a model accounts for or explains the location of the y values. The R^2 of a linear model describes how much smaller the variance (in the y direction) about the regression line is than the variance about the horizontal line $\bar{y}$. For example, consider the Elmhurst College data, shown in Figure 8.16. The variance of the response variable, aid received, is $s^2_{aid} = 29.8$. However, if we apply our least squares line, then this model reduces our uncertainty in predicting aid using a student's family income. The variability in the residuals describes how much variation remains after using the model: $s^2_{RES} = 22.4$. We could say that the reduction in the variance was:

$$\frac{s^2_{aid} - s^2_{RES}}{s^2_{aid}} = \frac{29.8 - 22.4}{29.8} = \frac{7.5}{29.8} = 0.25$$

If we used the simple standard deviation of the residuals, this would be exactly R^2. However, the standard way of computing the standard deviation of the residuals is slightly more sophisticated.[13] To avoid any trouble, we can instead use a sum of squares method. If we call the sum of the squared errors about the regression line $SSRes$ and the sum of the squared errors about the mean SSM, we can define R^2 as follows:

$$R^2 = \frac{SSM - SSRes}{SSM} = 1 - \frac{SSRes}{SSM}$$

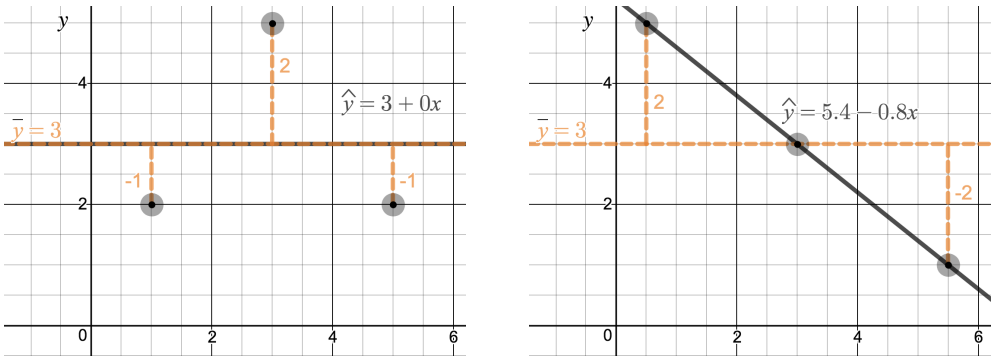


Figure 8.17: (a) The regression line is equivalent to $\bar{y}$; $R^2 = 0$. (b) The regression line passes through all of the points; $R^2 = 1$. Try out this and other interactive Desmos activities at openintro.org/ahss/desmos.

GUIDED PRACTICE 8.19

Using the formula for R^2, confirm that in Figure 8.17 (a), $R^2 = 0$ and that in Figure 8.17 (b), $R^2 = 1$.[14]

R^2 IS THE EXPLAINED VARIANCE

R^2 is always between 0 and 1, inclusive. It tells us the proportion of variation in the y values that is explained by a regression model. The higher the value of R^2, the better the model "explains" the response variable.

The value of R^2 is, in fact, equal to r^2, where r is the correlation. This means that $r = \pm\sqrt{R^2}$. Use this fact to answer the next two practice problems.

GUIDED PRACTICE 8.20

If a linear model has a very strong negative relationship with a correlation of -0.97, how much of the variation in the response variable is explained by the linear model?[15]

GUIDED PRACTICE 8.21

If a linear model has an R^2 or explained variance of 0.94, what is the correlation?[16]

[13] In computing the standard deviation of the residuals, we divide by $n-2$ rather than by $n-1$ to account for the $n-2$ degrees of freedom.

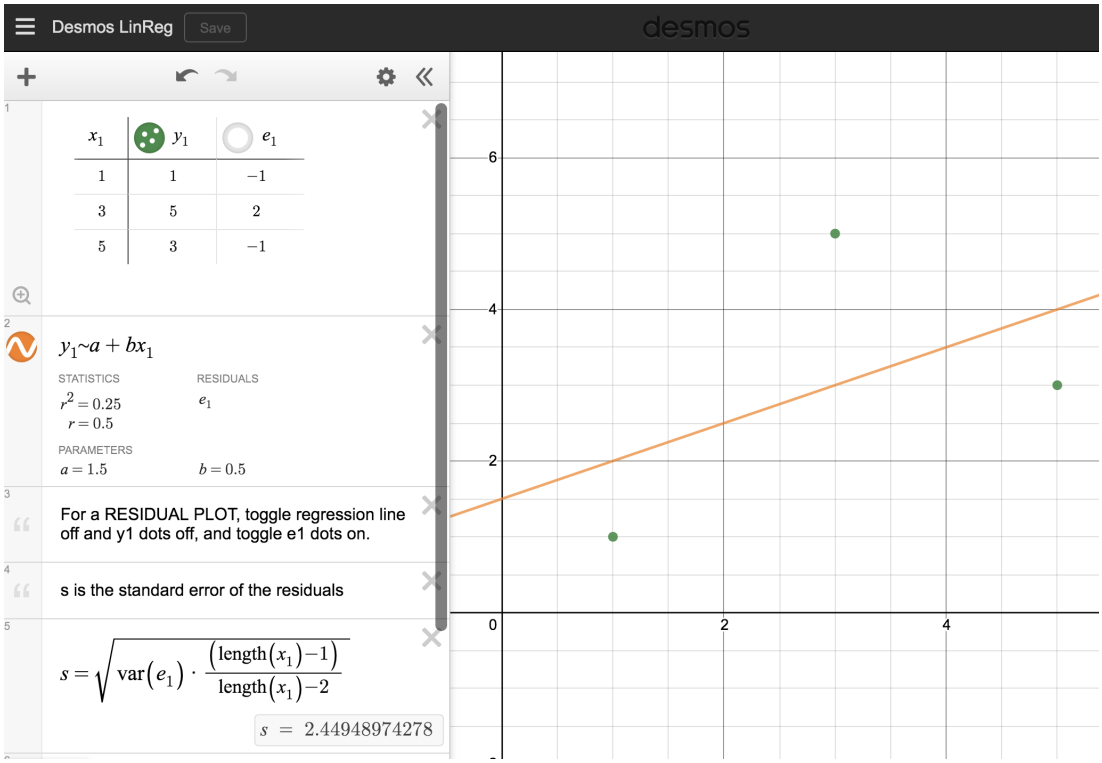
[14] (a) $SSRes = SSM = (-1)^2 + (2)^2 + (-1)^2 = 6$, so $R^2 = 1 - \frac{6}{6} = 0$. (b) $R^2 = 1 - \frac{0}{8}$.

[15] $R^2 = (-0.97)^2 = 0.94$ or 94%. 94% of the variation in y is explained by the linear model.

[16] We take the square root of R^2 and get 0.97, but we must be careful, because r could be 0.97 *or* -0.97. Without knowing the slope or seeing the scatterplot, we have no way of knowing if r is positive or negative.

8.2.6 Technology: linear correlation and regression

Get started quickly with this Desmos LinReg Calculator.



Calculator instructions

TI-84: FINDING a, b, R^2, AND r FOR A LINEAR MODEL

Use `STAT`, `CALC`, `LinReg(a + bx)`.

1. Choose `STAT`.
2. Right arrow to `CALC`.
3. Down arrow and choose `8:LinReg(a+bx)`.
   - Caution: choosing `4:LinReg(ax+b)` will reverse a and b.
4. Let `Xlist` be `L1` and `Ylist` be `L2` (don't forget to enter the x and y values in L1 and L2 before doing this calculation).
5. Leave `FreqList` blank.
6. Leave `Store RegEQ` blank.
7. Choose Calculate and hit `ENTER`, which returns:

`a`	a, the y-intercept of the best fit line
`b`	b, the slope of the best fit line
`r²`	R^2, the explained variance
`r`	r, the correlation coefficient

TI-83: Do steps 1-3, then enter the x list and y list separated by a comma, e.g. `LinReg(a+bx) L1, L2`, then hit `ENTER`.

WHAT TO DO IF R^2 AND r DO NOT SHOW UP ON A TI-83/84

If r^2 and r do now show up when doing STAT, CALC, LinReg, the *diagnostics* must be turned on. This only needs to be once and the diagnostics will remain on.

1. Hit 2ND 0 (i.e. CATALOG).
2. Scroll down until the arrow points at DiagnosticOn.
3. Hit ENTER and ENTER again. The screen should now say:

```
DiagnosticOn
               Done
```

WHAT TO DO IF A TI-83/84 RETURNS: ERR: DIM MISMATCH

This error means that the lists, generally L1 and L2, do not have the same length.

1. Choose 1:Quit.
2. Choose STAT, Edit and make sure that the lists have the same number of entries.

CASIO FX-9750GII: FINDING a, b, R^2, AND r FOR A LINEAR MODEL

1. Navigate to STAT (MENU button, then hit the 2 button or select STAT).
2. Enter the x and y data into 2 separate lists, e.g. x values in List 1 and y values in List 2. Observation ordering should be the same in the two lists. For example, if $(5, 4)$ is the second observation, then the second value in the x list should be 5 and the second value in the y list should be 4.
3. Navigate to CALC (F2) and then SET (F6) to set the regression context.
   - To change the 2Var XList, navigate to it, select List (F1), and enter the proper list number. Similarly, set 2Var YList to the proper list.
4. Hit EXIT.
5. Select REG (F3), X (F1), and a+bx (F2), which returns:

a	a, the y-intercept of the best fit line
b	b, the slope of the best fit line
r	r, the correlation coefficient
r^2	R^2, the explained variance
MSe	Mean squared error, which you can ignore

   If you select ax+b (F1), the a and b meanings will be reversed.

GUIDED PRACTICE 8.22

The data set `loan50`, introduced in Chapter 1, contains information on randomly sampled loans offered through Lending Club. A subset of the data matrix is shown in Figure 8.18. Use a calculator to find the equation of the least squares regression line for predicting loan amount from total income based on this subset.[17]

	total_income	loan_amount
1	59000	22000
2	60000	6000
3	75000	25000
4	75000	6000
5	254000	25000
6	67000	6400
7	28800	3000

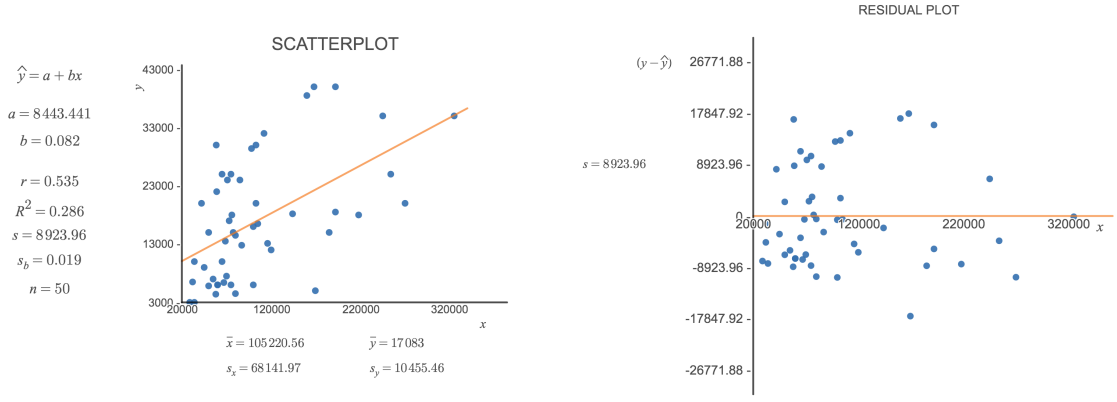
Figure 8.18: Sample of data from `loan50`.

GUIDED PRACTICE 8.23

Use the loan50 data set and Screen 2 of this Desmos LinReg Calculator to draw the scatterplot and find the equation of the least squares regression line for prediction loan amount (y) from total income (x). Also draw the residual plot and find the standard error of the residuals.[18]

[17] a = 11121 and b = 0.0043, therefore $\hat{y} = 11121 + 0.0043x$.

[18] Recall that data sets can be found at openintro.org/data and Desmos Calculators can be found at openintro.org/ahss/desmos. This will first require a minor data cleaning step. After downloading the loan50 data, notice the oddly formatted numbers under total_income column. This will confuse Desmos! Highight the column, do Format, Cells, Number, and set decimal places to 0. Finally, copy and paste the loan_amount column to the right of total_income. Then you can copy the two columns into Desmos.



8.2.7 Types of outliers in linear regression

Outliers in regression are observations that fall far from the "cloud" of points. These points are especially important because they can have a strong influence on the least squares line.

EXAMPLE 8.24

There are six plots shown in Figure 8.19 along with the least squares line and residual plots. For each scatterplot and residual plot pair, identify any obvious outliers and note how they influence the least squares line. Recall that an outlier is any point that doesn't appear to belong with the vast majority of the other points.

(1) There is one outlier far from the other points, though it only appears to slightly influence the line.

(2) There is one outlier on the right, though it is quite close to the least squares line, which suggests it wasn't very influential.

(E) (3) There is one point far away from the cloud, and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud doesn't appear to fit very well.

(4) There is a primary cloud and then a small secondary cloud of four outliers. The secondary cloud appears to be influencing the line somewhat strongly, making the least squares line fit poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated.

(5) There is no obvious trend in the main cloud of points and the outlier on the right appears to largely control the slope of the least squares line.

(6) There is one outlier far from the cloud, however, it falls quite close to the least squares line and does not appear to be very influential.

Examine the residual plots in Figure 8.19. You will probably find that there is some trend in the main clouds of (3) and (4). In these cases, the outliers influenced the slope of the least squares lines. In (5), data with no clear trend were assigned a line with a large trend simply due to one outlier (!).

LEVERAGE

Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with **high leverage**.

Points that fall horizontally far from the line are points of high leverage; these points can strongly influence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line – as in cases (3), (4), and (5) of Example 8.24 – then we call it an **influential point**. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line.

It is tempting to remove outliers. Don't do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly. For instance, if a financial firm ignored the largest market swings – the "outliers" – they would soon go bankrupt by making poorly thought-out investments.

DON'T IGNORE OUTLIERS WHEN FITTING A FINAL MODEL

If there are outliers in the data, they should not be removed or ignored without a good reason. Whatever final model is fit to the data would not be very helpful if it ignores the most exceptional cases.

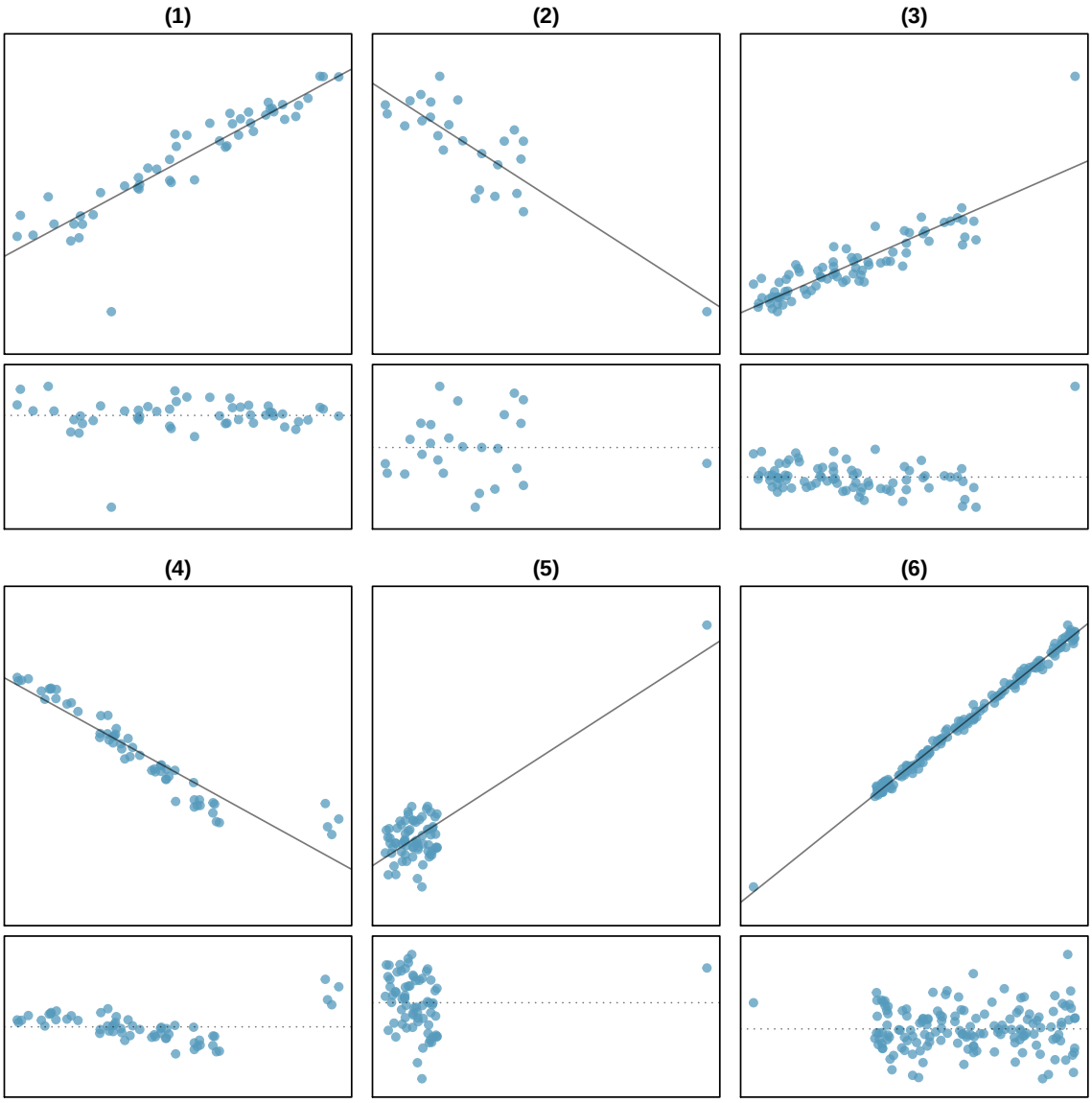


Figure 8.19: Six plots, each with a least squares line and residual plot. All data sets have at least one outlier.

8.2.8 Categorical predictors with two levels (special topic)

Categorical variables are also useful in predicting outcomes. Here we consider a categorical predictor with two levels (recall that a *level* is the same as a *category*). We'll consider eBay auctions for a video game, *Mario Kart* for the Nintendo Wii, where both the total price of the auction and the condition of the game were recorded.[19] Here we want to predict total price based on game condition, which takes values `used` and `new`. A plot of the auction data is shown in Figure 8.20.

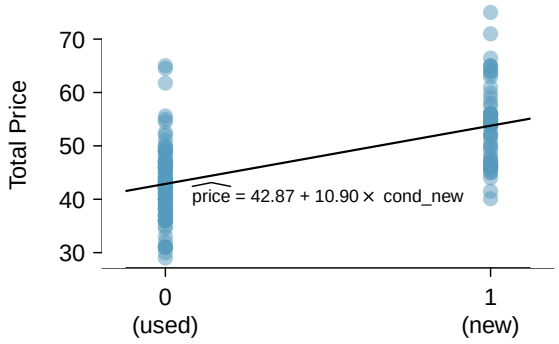


Figure 8.20: Total auction prices for the game *Mario Kart*, divided into used ($x = 0$) and new ($x = 1$) condition games with the least squares regression line shown.

To incorporate the game condition variable into a regression equation, we must convert the categories into a numerical form. We will do so using an **indicator variable** called `cond_new`, which takes value 1 when the game is new and 0 when the game is used. Using this indicator variable, the linear model may be written as

$$\widehat{price} = \alpha + \beta \times \texttt{cond\_new}$$

The fitted model is summarized in Figure 8.21, and the model with its parameter estimates is given as

$$\widehat{price} = 42.87 + 10.90 \times \texttt{cond\_new}$$

For categorical predictors with two levels, the linearity assumption will always be satisfied. However, we must evaluate whether the residuals in each group are approximately normal with equal variance. Based on Figure 8.20, both of these conditions are reasonably satisfied.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.87	0.81	52.67	0.0000
cond_new	10.90	1.26	8.66	0.0000

Figure 8.21: Least squares regression summary for the Mario Kart data.

EXAMPLE 8.25

Interpret the two parameters estimated in the model for the price of *Mario Kart* in eBay auctions.

E

The intercept is the estimated price when `cond_new` takes value 0, i.e. when the game is in used condition. That is, the average selling price of a used version of the game is $42.87.

The slope indicates that, on average, new games sell for about $10.90 more than used games.

INTERPRETING MODEL ESTIMATES FOR CATEGORICAL PREDICTORS.

The estimated intercept is the value of the response variable for the first category (i.e. the category corresponding to an indicator value of 0). The estimated slope is the average change in the response variable between the two categories.

[19] These data were collected in Fall 2009 and may be found at openintro.org/stat.

Section summary

- We define the *best fit line* as the line that minimizes the sum of the squared residuals (errors) about the line. That is, we find the line that minimizes $(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+\cdots+(y_n-\hat{y}_n)^2 = \sum (y_i-\hat{y}_i)^2$. We call this line the **least squares regression line**.
- We write the least squares regression line in the form: $\hat{y} = a + bx$, and we can calculate a and b based on the summary statistics as follows:

$$b = r\frac{s_y}{s_x} \qquad \text{and} \qquad a = \bar{y} - b\bar{x}.$$

- *Interpreting* the **slope** and **y-intercept** of a linear model
  - The slope, b, describes the *average* increase or decrease in the y variable if the explanatory variable x is one unit larger.
  - The y-intercept, a, describes the average or predicted outcome of y if $x = 0$. The linear model must be valid all the way to $x = 0$ for this to make sense, which in many applications is not the case.
- Two important considerations about the regression line
  - The regression line provides *estimates* or *predictions*, not actual values. It is important to know how large s, the standard deviation of the residuals, is in order to know about how much error to expect in these predictions.
  - The regression line estimates are only reasonable within the domain of the data. Predicting y for x values that are outside the domain, known as **extrapolation**, is unreliable and may produce ridiculous results.
- Using R^2 to assess the fit of the model
  - R^2, called **R-squared** or the **explained variance**, is a measure of how well the model explains or fits the data. R^2 is always between 0 and 1, inclusive, or between 0% and 100%, inclusive. The higher the value of R^2, the better the model "fits" the data.
  - The R^2 for a linear model describes the *proportion of variation* in the y variable that is *explained by* the regression line.
  - R^2 applies to any type of model, not just a linear model, and can be used to compare the fit among various models.
  - The correlation $r = -\sqrt{R^2}$ or $r = \sqrt{R^2}$. The value of R^2 is always positive and cannot tell us the *direction* of the association. If finding r based on R^2, make sure to use either the scatterplot or the slope of the regression line to determine the *sign* of r.
- When a residual plot of the data appears as a random cloud of points, a linear model is generally appropriate. If a residual plot of the data has any type of pattern or curvature, such as a $\cup$-shape, a linear model is not appropriate.
- **Outliers** in regression are observations that fall far from the "cloud" of points.
- An **influential point** is a point that has a big effect or pull on the slope of the regression line. Points that are outliers in the x direction will have more pull on the slope of the regression line and are more likely to be influential points.

Exercises

8.17 Units of regression. Consider a regression predicting weight (kg) from height (cm) for a sample of adult males. What are the units of the correlation coefficient, the intercept, and the slope?

8.18 Which is higher? Determine if I or II is higher or if they are equal. Explain your reasoning. For a regression line, the uncertainty associated with the slope estimate, b_1, is higher when

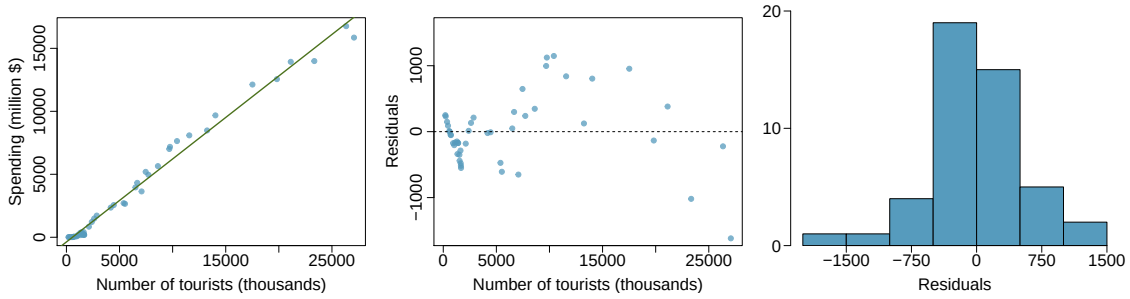
I. there is a lot of scatter around the regression line or

II. there is very little scatter around the regression line

8.19 Over-under, Part I. Suppose we fit a regression line to predict the shelf life of an apple based on its weight. For a particular apple, we predict the shelf life to be 4.6 days. The apple's residual is -0.6 days. Did we over or under estimate the shelf-life of the apple? Explain your reasoning.

8.20 Over-under, Part II. Suppose we fit a regression line to predict the number of incidents of skin cancer per 1,000 people from the number of sunny days in a year. For a particular year, we predict the incidence of skin cancer to be 1.5 per 1,000 people, and the residual for this year is 0.5. Did we over or under estimate the incidence of skin cancer? Explain your reasoning.

8.21 Tourism spending. The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year.[20] Three plots are provided: scatterplot showing the relationship between these two variables along with the least squares fit, residuals plot, and histogram of residuals.



(a) Describe the relationship between number of tourists and spending.

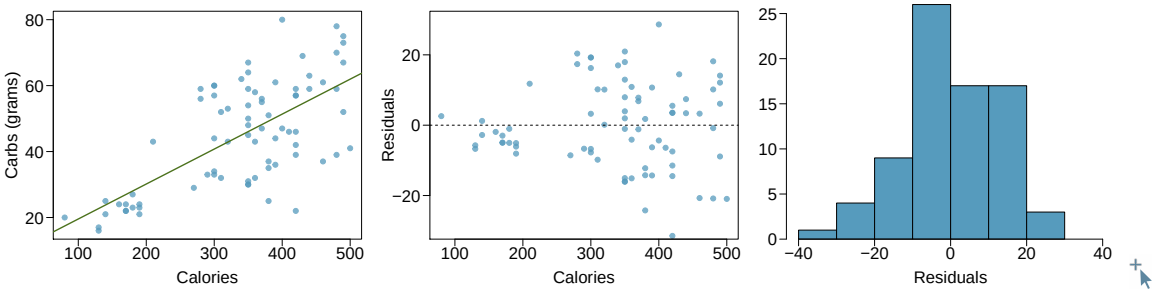
(b) What are the explanatory and response variables?

(c) Why might we want to fit a regression line to these data?

(d) Do the data meet the conditions required for fitting a least squares line? In addition to the scatterplot, use the residual plot and histogram to answer this question.

[20] **data:turkeyTourism**.

8.22 Nutrition at Starbucks, Part I. The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.[21] Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content.



(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

(b) In this scenario, what are the explanatory and response variables?

(c) Why might we want to fit a regression line to these data?

(d) Do these data meet the conditions required for fitting a least squares line?

8.23 The Coast Starlight, Part II. Exercise 8.11 introduces data on the Coast Starlight Amtrak train that runs from Seattle to Los Angeles. The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

(a) Write the equation of the regression line for predicting travel time.

(b) Interpret the slope and the intercept in this context.

(c) Calculate R^2 of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret R^2 in the context of the application.

(d) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.

(e) It actually takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.

(f) Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

8.24 Body measurements, Part III. Exercise 8.13 introduces data on shoulder girth and height of a group of individuals. The mean shoulder girth is 107.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67.

(a) Write the equation of the regression line for predicting height.

(b) Interpret the slope and the intercept in this context.

(c) Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

(d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.

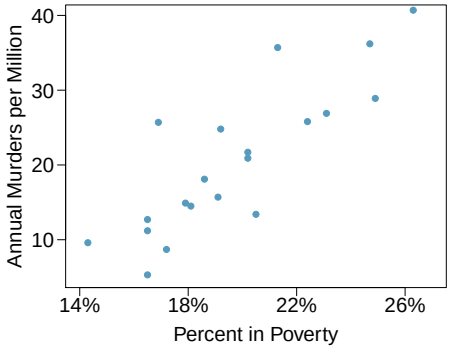
(e) The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.

(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

[21] **data:starbucksCals**.

8.25 Murders and poverty, Part I. The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

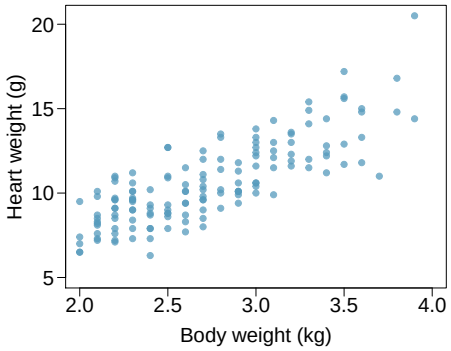
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-29.901	7.789	-3.839	0.001	$s =$
poverty%	2.559	0.390	6.562	0.000	
5.512	$R^2 = 70.52\%$			$R^2_{adj} = 68.89\%$	



(a) Write out the linear model.
(b) Interpret the intercept.
(c) Interpret the slope.
(d) Interpret R^2.
(e) Calculate the correlation coefficient.

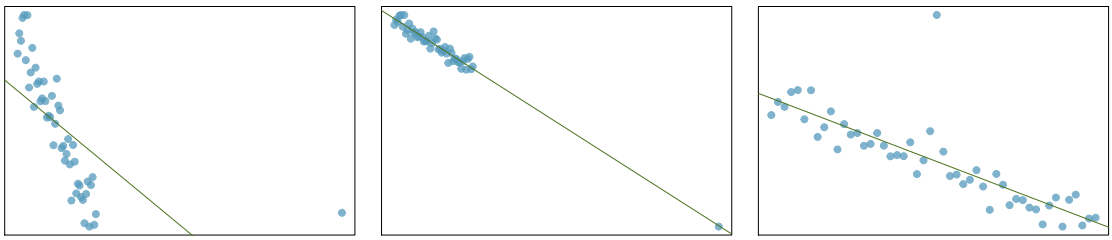
8.26 Cats, Part I. The following regression output is for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.357	0.692	-0.515	0.607	$s =$
body wt	4.034	0.250	16.119	0.000	
1.452	$R^2 = 64.66\%$			$R^2_{adj} = 64.41\%$	

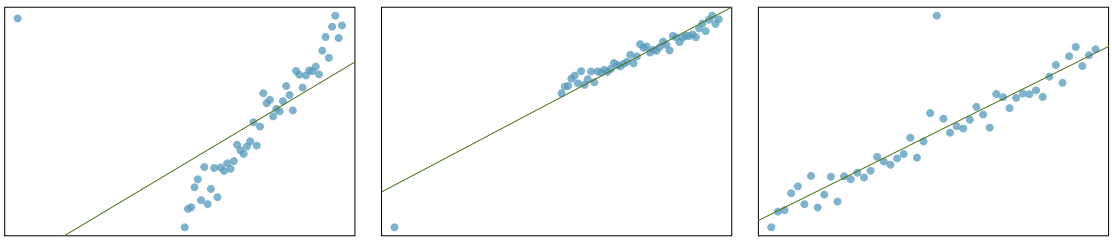


(a) Write out the linear model.
(b) Interpret the intercept.
(c) Interpret the slope.
(d) Interpret R^2.
(e) Calculate the correlation coefficient.

8.27 Outliers, Part I. Identify the outliers in the scatterplots shown below, and determine what type of outliers they are. Explain your reasoning.



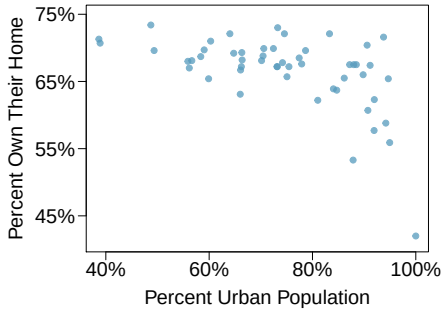
8.28 Outliers, Part II. Identify the outliers in the scatterplots shown below and determine what type of outliers they are. Explain your reasoning.



8.29 Urban homeowners, Part I. The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas.[22] There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.

(a) Describe the relationship between the percent of families who own their home and the percent of the population living in urban areas.

(b) The outlier at the bottom right corner is District of Columbia, where 100% of the population is considered urban. What type of an outlier is this observation?



8.30 Crawling babies, Part II. Exercise 8.12 introduces data on the average monthly temperature during the month babies first try to crawl (about 6 months after birth) and the average first crawling age for babies born in a given month. A scatterplot of these two variables reveals a potential outlying month when the average temperature is about 53°F and average crawling age is about 28.5 weeks. Does this point have high leverage? Is it an influential point?

[22] **data:urbanOwner**.

8.3 Transformations for skewed data

County population size among the counties in the US is very strongly right skewed. Can we apply a transformation to make the distribution more symmetric? How would such a transformation affect the scatterplot and residual plot when another variable is graphed against this variable? In this section, we will see the power of transformations for very skewed data.

Learning objectives

1. See how a log transformation can bring symmetry to an extremely skewed variable.
2. Recognize that data can often be transformed to produce a linear relationship, and that this transformation often involves log of the y-values and sometimes log of the x-values.
3. Use residual plots to assess whether a linear model for transformed data is reasonable.

8.3.1 Introduction to transformations

EXAMPLE 8.26

Consider the histogram of county populations shown in Figure 8.22(a), which shows extreme skew. What isn't useful about this plot?

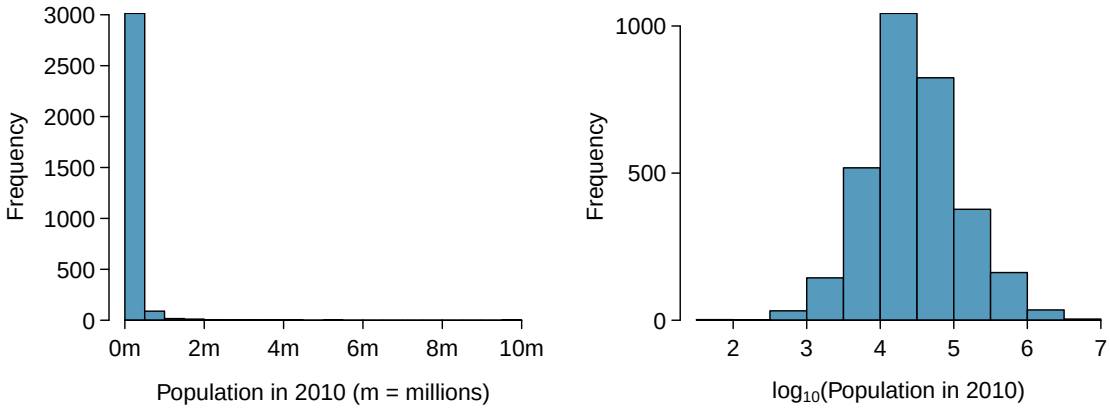
(E)

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

There are some standard transformations that may be useful for strongly right skewed data where much of the data is positive but clustered near zero. A **transformation** is a rescaling of the data using a function. For instance, a plot of the logarithm (base 10) of county populations results in the new histogram in Figure 8.22(b). This data is symmetric, and any potential outliers appear much less extreme than in the original data set. By reigning in the outliers and extreme skew, transformations like this often make it easier to build statistical models against the data.

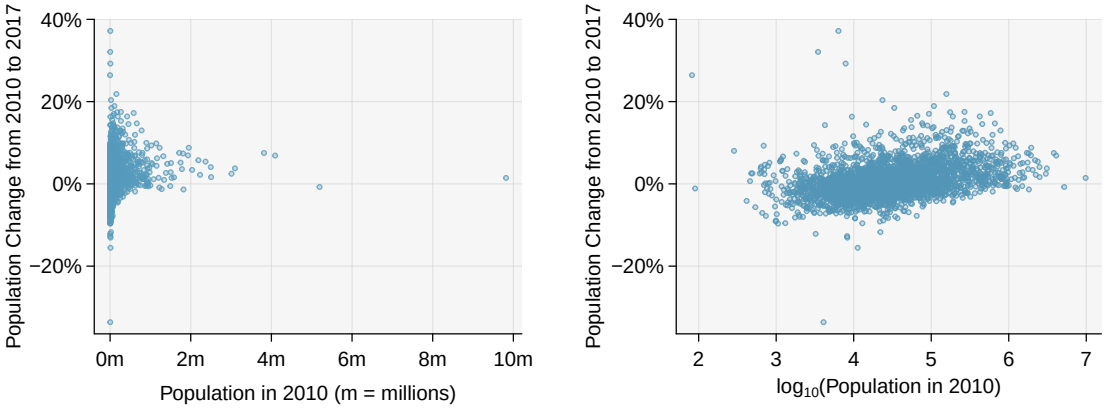
Transformations can also be applied to one or both variables in a scatterplot. A scatterplot of the population change from 2010 to 2017 against the population in 2010 is shown in Figure 8.23(a). In this first scatterplot, it's hard to decipher any interesting patterns because the population variable is so strongly skewed. However, if we apply a $\log_{10}$ transformation to the population variable, as shown in Figure 8.23(b), a positive association between the variables is revealed. While fitting a line to predict population change (2010 to 2017) from population (in 2010) does not seem reasonable, fitting a line to predict population from $\log_{10}$(population) does seem reasonable.

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



(a) (b)

Figure 8.22: (a) A histogram of the populations of all US counties. (b) A histogram of $\log_{10}$-transformed county populations. For this plot, the x-value corresponds to the power of 10, e.g. "4" on the x-axis corresponds to $10^4 = 10{,}000$.



(a) (b)

Figure 8.23: (a) Scatterplot of population change against the population before the change. (b) A scatterplot of the same data but where the population size has been log-transformed.

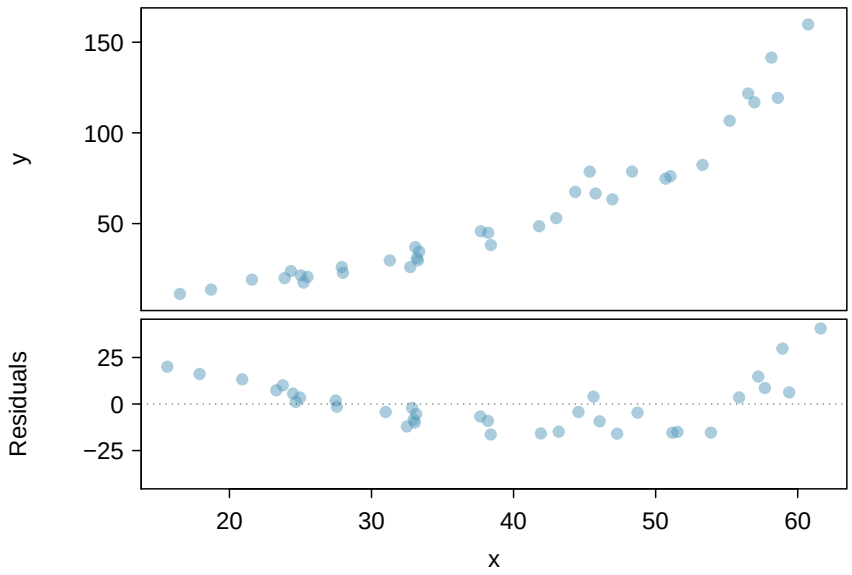


Figure 8.24: Variable y is plotted against x. A nonlinear relationship is evident by the ∪-pattern shown in the residual plot. The curvature is also visible in the original plot.

8.3.2 Transformations to achieve linearity

EXAMPLE 8.27

Consider the scatterplot and residual plot in Figure 8.24. The regression output is also provided. Is the linear model $\hat{y} = -52.3564 + 2.7842x$ a good model for the data?

```
The regression equation is

y = -52.3564 + 2.7842 x

Predictor       Coef   SE Coef         T          P
Constant    -52.3564    7.2757    -7.196      3e-08
x             2.7842    0.1768    15.752    < 2e-16

S = 13.76    R-Sq = 88.26%    R-Sq(adj) = 87.91%
```

(E)

We can note the R^2 value is fairly large. However, this alone does not mean that the model is good. Another model might be much better. When assessing the appropriateness of a linear model, we should look at the residual plot. The ∪-pattern in the residual plot tells us the original data is curved. If we inspect the two plots, we can see that for small and large values of x we systematically underestimate y, whereas for middle values of x, we systematically overestimate y. The curved trend can also be seen in the original scatterplot. Because of this, the linear model is not appropriate, and it would not be appropriate to perform a t-test for the slope because the conditions for inference are not met. However, we might be able to use a transformation to linearize the data.

Regression analysis is easier to perform on linear data. When data are nonlinear, we sometimes **transform** the data in a way that makes the resulting relationship linear. The most common **transformation** is log of the y values. Sometimes we also apply a transformation to the x values. We generally use the residuals as a way to evaluate whether the transformed data are more linear. If so, we can say that a better model has been found.

EXAMPLE 8.28

Using the regression output for the transformed data, write the new linear regression equation.

E

```
The regression equation is

log(y) = 1.722540 + 0.052985 x

Predictor         Coef     SE Coef        T          P
Constant      1.722540    0.056731    30.36    < 2e-16
x             0.052985    0.001378    38.45    < 2e-16

S = 0.1073    R-Sq = 97.82%    R-Sq(adj) = 97.75%
```

The linear regression equation can be written as: $\widehat{\log(y)} = 1.723 + 0.053x$

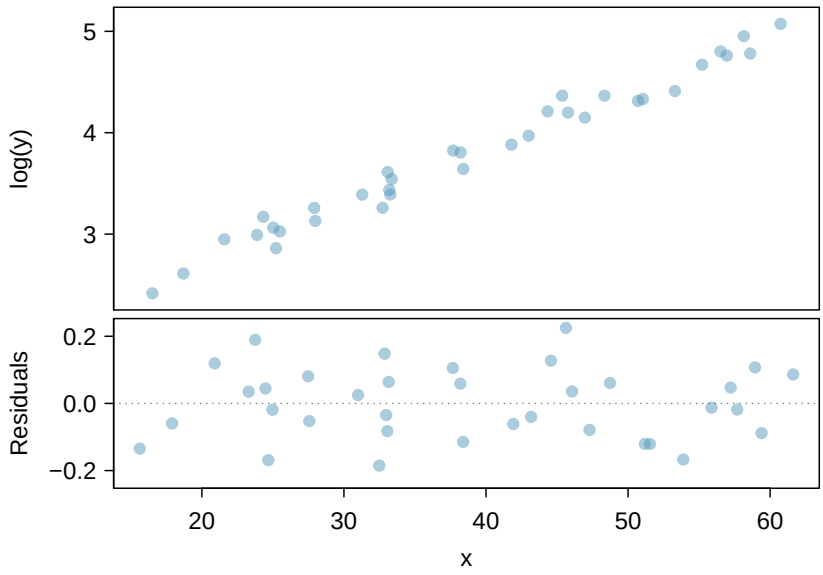


Figure 8.25: A plot of $\log(y)$ against x. The residuals don't show any evident patterns, which suggests the transformed data is well-fit by a linear model.

GUIDED PRACTICE 8.29

Which of the following statements are true? There may be more than one.[23]

G

(a) There is an apparent linear relationship between x and y.

(b) There is an apparent linear relationship between x and $\widehat{\log(y)}$.

(c) The model provided by Regression I ($\hat{y} = -52.3564 + 2.7842x$) yields a better fit.

(d) The model provided by Regression II ($\widehat{\log(y)} = 1.723 + 0.053x$) yields a better fit.

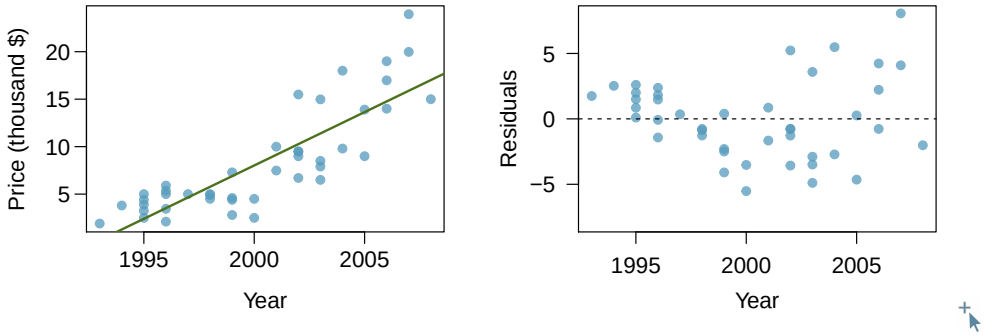
[23] Part (a) is *false* since there is a nonlinear (curved) trend in the data. Part (b) is true. Since the transformed data shows a stronger linear trend, it is a better fit, i.e. Part (c) is *false*, and Part (d) is true.

Section summary

- A **transformation** is a rescaling of the data using a function. When data are very skewed, a log transformation often results in more symmetric data.
- Regression analysis is easier to perform on linear data. When data are nonlinear, we sometimes **transform** the data in a way that results in a linear relationship. The most common transformation is log of the y-values. Sometimes we also apply a transformation to the x-values.
- To assess the model, we look at the **residual plot** of the *transformed* data. If the residual plot of the original data has a pattern, but the residual plot of the transformed data has no pattern, a linear model for the transformed data is reasonable, and the transformed model provides a better fit than the simple linear model.

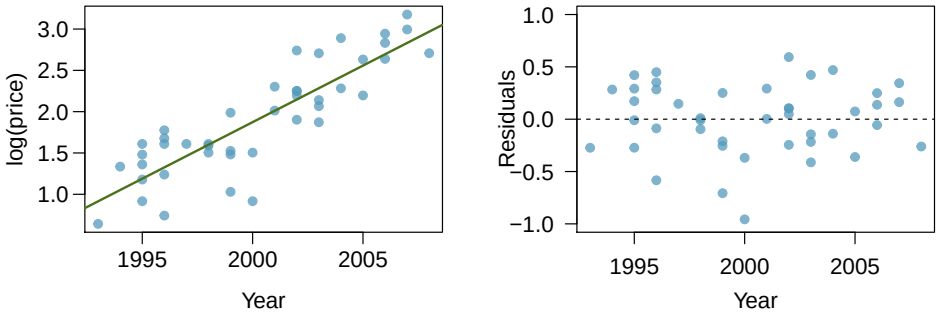
Exercises

8.31 Used trucks. The scatterplot below shows the relationship between year and price (in thousands of $) of a random sample of 42 pickup trucks. Also shown is a residuals plot for the linear model for predicting price from year.



(a) Describe the relationship between these two variables and comment on whether a linear model is appropriate for modeling the relationship between year and price.

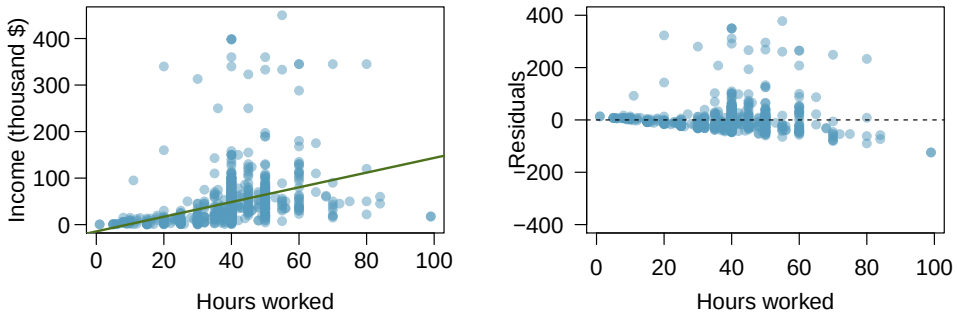
(b) The scatterplot below shows the relationship between logged (natural log) price and year of these trucks, as well as the residuals plot for modeling these data. Comment on which model (linear model from earlier or logged model presented here) is a better fit for these data.



(c) The output for the logged model is given below. Interpret the slope in context of the data.

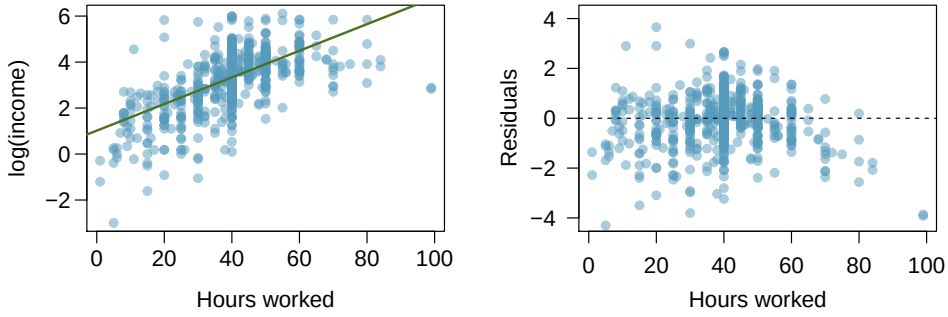
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-271.981	25.042	-10.861	0.000
Year	0.137	0.013	10.937	0.000

8.32 Income and hours worked. The scatterplot below shows the relationship between income and years worked for a random sample of 787 Americans. Also shown is a residuals plot for the linear model for predicting income from hours worked. The data come from the 2012 American Community Survey.[24]



(a) Describe the relationship between these two variables and comment on whether a linear model is appropriate for modeling the relationship between year and price.

(b) The scatterplot below shows the relationship between logged (natural log) income and hours worked, as well as the residuals plot for modeling these data. Comment on which model (linear model from earlier or logged model presented here) is a better fit for these data.



(c) The output for the logged model is given below. Interpret the slope in context of the data.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.017	0.113	9.000	0.000
hrs_work	0.058	0.003	21.086	0.000

[24] **data:acs:2012**.

8.4 Inference for the slope of a regression line

Here we encounter our last confidence interval and hypothesis test procedures, this time for making inferences about the slope of the population regression line. We can use this to answer questions such as the following:

- Is the unemployment rate a significant linear predictor for the loss of the President's party in the House of Representatives?
- On average, how much less in college gift aid do students receive when their parents earn an additional $1000 in income?

Learning objectives

1. Recognize that the slope of the sample regression line is a point estimate and has an associated standard error.
2. Be able to read the results of computer regression output and identify the quantities needed for inference for the slope of the regression line, specifically the slope of the sample regression line, the SE of the slope, and the degrees of freedom.
3. State and verify whether or not the conditions are met for inference on the slope of the regression line based using the t-distribution.
4. Carry out a complete confidence interval procedure for the slope of the regression line.
5. Carry out a complete hypothesis test for the slope of the regression line.
6. Distinguish between when to use the t-test for the slope of a regression line and when to use the paired t-test for a mean of differences.

8.4.1 The role of inference for regression parameters

Previously, we found the equation of the regression line for predicting gift aid from family income at Elmhurst College. The slope, b, was equal to -0.0431. This is the slope for our sample data. However, the sample was taken from a larger population. We would like to use the slope computed from our sample data to estimate the slope of the population regression line.

The equation for the population regression line can be written as

$$\mu_y = \alpha + \beta x$$

Here, α and β represent two model parameters, namely the y-intercept and the slope of the true or population regression line. (This use of α and β have nothing to do with the α and β we used previously to represent the probability of a Type I Error and Type II Error!) The parameters α and β are estimated using data. We can look at the equation of the regression line calculated from a particular data set:

$$\hat{y} = a + bx$$

and see that a and b are point estimates for α and β, respectively. If we plug in the values of a and b, the regression equation for predicting gift aid based on family income is:

$$\hat{y} = 24.3193 - 0.0431x$$

The slope of the sample regression line, -0.0431, is our best estimate for the slope of the population regression line, but there is variability in this estimate since it is based on a sample. A different sample would produce a somewhat different estimate of the slope. The standard error of the slope tells us the typical variation in the slope of the sample regression line and the typical error in using this slope to estimate the slope of the population regression line.

We would like to construct a 95% confidence interval for β, the slope of the population regression line. As with means, inference for the slope of a regression line is based on the t-distribution.

INFERENCE FOR THE SLOPE OF A REGRESSION LINE

Inference for the slope of a regression line is based on the t-distribution with $n - 2$ degrees of freedom, where n is the number of paired observations.

Once we verify that conditions for using the t-distribution are met, we will be able to construct the confidence interval for the slope using a critical value $t^\star$ based on $n - 2$ degrees of freedom. We will use a table of the regression summary to find the point estimate and standard error for the slope.

8.4.2 Conditions for the least squares line

Conditions for inference in the context of regression can be more complicated than when dealing with means or proportions.

Inference for parameters of a regression line involves the following assumptions:

Linearity. The true relationship between the two variables follows a linear trend. We check whether this is reasonable by examining whether the data follows a linear trend. If there is a nonlinear trend (e.g. left panel of Figure 8.26), an advanced regression method from another book or later course should be applied.

Nearly normal residuals. For each x-value, the residuals should be nearly normal. When this assumption is found to be unreasonable, it is usually because of outliers or concerns about influential points. An example which suggestions non-normal residuals is shown in the second panel of Figure 8.26. If the sample size $n \geq 30$, then this assumption is not necessary.

Constant variability. The variability of points around the true least squares line is constant for all values of x. An example of non-constant variability is shown in the third panel of Figure 8.26.

Independent. The observations are independent of one other. The observations can be considered independent when they are collected from a random sample or randomized experiment. Be careful of data collected sequentially in what is called a **time series**. An example of data collected in such a fashion is shown in the fourth panel of Figure 8.26.

We see in Figure 8.26, that patterns in the residual plots suggest that the assumptions for regression inference are not met in those four examples. In fact, identifying nonlinear trends in the data, outliers, and non-constant variability in the residuals are often easier to detect in a residual plot than in a scatterplot.

We note that the second assumption regarding nearly normal residuals is particularly difficult to assess when the sample size is small. We can make a graph, such as a histogram, of the residuals, but we cannot expect a small data set to be nearly normal. All we can do is to look for excessive skew or outliers. Outliers and influential points in the data can be seen from the residual plot as well as from a histogram of the residuals.

CONDITIONS FOR INFERENCE ON THE SLOPE OF A REGRESSION LINE

1. The data is collected from a random sample or randomized experiment.
2. The residual plot appears as a random cloud of points and does not have any patterns or significant outliers that would suggest that the linearity, nearly normal residuals, constant variability, or independence assumptions are unreasonable.

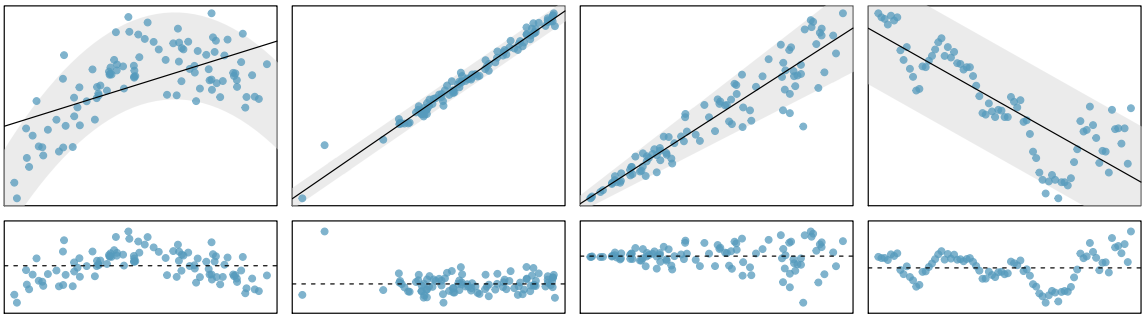


Figure 8.26: Four examples showing when the inference methods in this chapter are insufficient to apply to the data. In the left panel, a straight line does not fit the data. In the second panel, there are outliers; two points on the left are relatively distant from the rest of the data, and one of these points is very far away from the line. In the third panel, the variability of the data around the line increases with larger values of x. In the last panel, a time series data set is shown, where successive observations are highly correlated.

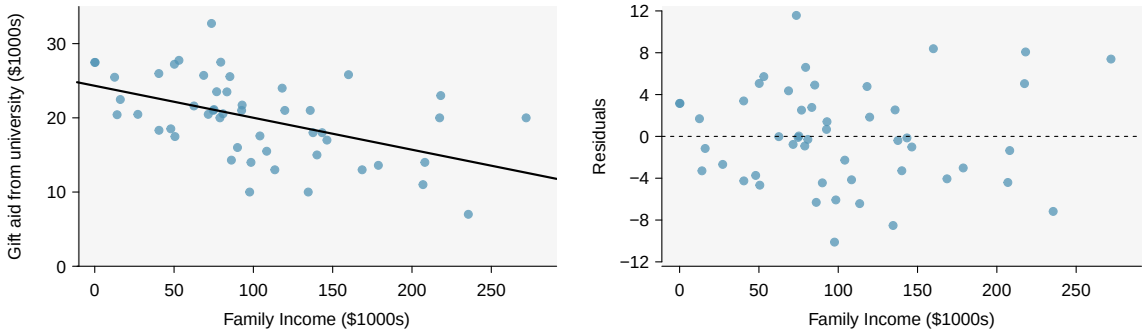


Figure 8.27: Left: Scatterplot of gift aid versus family income for 50 freshmen at Elmhurst college. Right: Residual plot for the model shown in left panel.

8.4.3 Constructing a confidence interval for the slope of a regression line

We would like to construct a confidence interval for the slope of the regression line for predicting gift aid based on family income for *all* freshmen at Elmhurst college.

Do conditions seem to be satisfied? We recall that the 50 freshmen in the sample were randomly chosen, so the observations are independent. Next, we need to look carefully at the scatterplot and the residual plot.

ALWAYS CHECK CONDITIONS

Do not blindly apply formulas or rely on regression output; always first look at a scatterplot or a residual plot. If conditions for fitting the regression line are not met, the methods presented here should not be applied.

The scatterplot seems to show a linear trend, which matches the fact that there is no curved trend apparent in the residual plot. Also, the standard deviation of the residuals is mostly constant for different x values and there are no outliers or influential points. There are no patterns in the residual plot that would suggest that a linear model is not appropriate, so the conditions are reasonably met. We are now ready to calculate the 95% confidence interval.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.3193	1.2915	18.83	0.0000
family_income	-0.0431	0.0108	-3.98	0.0002

Figure 8.28: Summary of least squares fit for the Elmhurst College data, where we are predicting gift aid by the university based on the family income of students.

EXAMPLE 8.30

Construct a 95% confidence interval for the slope of the regression line for predicting gift aid from family income at Elmhurst college.

As usual, the confidence interval will take the form:

$$\text{point estimate} \pm \text{critical value} \times SE \text{ of estimate}$$

(E) The point estimate for the slope of the population regression line is the slope of the sample regression line: -0.0431. The standard error of the slope can be read from the table as 0.0108. Note that we do not need to divide 0.0108 by the square root of n or do any further calculations on 0.0108; 0.0108 *is* the SE of the slope. Note that the value of t given in the table refers to the test statistic, not to the critical value $t^\star$. To find $t^\star$ we can use a t-table. Here $n = 50$, so $df = 50 - 2 = 48$. Using a t-table, we round down to row $df = 40$ and we estimate the critical value $t^\star = 2.021$ for a 95% confidence level. The confidence interval is calculated as:

$$\begin{aligned} &-0.0431 \pm 2.021 \times 0.0108 \\ &= (-0.065, -0.021) \end{aligned}$$

Note: $t^\star$ using exactly 48 degrees of freedom is equal to 2.01 and gives the same interval of $(-0.065,\ -0.021)$.

EXAMPLE 8.31

Intepret the confidence interval in context. What can we conclude?

(E) We are 95% confident that the slope of the population regression line, the true average change in gift aid for each additional \$1000 in family income, is between $-$\$0.065 thousand dollars and $-$\$0.021 thousand dollars. That is, we are 95% confident that, on average, when family income is \$1000 higher, gift aid is between \$21 and \$65 *lower*.

Because the entire interval is negative, we have evidence that the slope of the population regression line is less than 0. In other words, we have evidence that there is a significant negative linear relationship between gift aid and family income.

CONSTRUCTING A CONFIDENCE INTERVAL FOR THE SLOPE OF REGRESSION LINE

To carry out a complete confidence interval procedure to estimate the slope of the population regression line β,

Identify: Identify the parameter and the confidence level, C%.

The parameter will be a slope of the population regression line, e.g. the slope of the population regression line relating air quality index to average rainfall per year for each city in the United States.

Choose: Choose the correct interval procedure and identify it by name.

To estimate the slope of a regression model we use a ***t*-interval for the slope**.

Check: Check conditions for using a t-interval for the slope.

1. Independence: Data should come from a random sample or randomized experiment. If sampling without replacement, check that the sample size is less than 10% of the population size.
2. Linearity: Check that the scatterplot does not show a curved trend and that the residual plot shows no $\cup$-shape pattern.
3. Constant variability: Use the residual plot to check that the standard deviation of the residuals is constant across all x-values.
4. Normality: The population of residuals is nearly normal or the sample size is ≥ 30. If the sample size is less than 30 check for strong skew or outliers in the sample residuals. If neither is found, then the condition that the population of residuals is nearly normal is considered reasonable.

Calculate: Calculate the confidence interval and record it in interval form.

point estimate $\pm$ $t^\star \times SE$ of estimate, $df = n - 2$

point estimate: the slope b of the sample regression line

SE of estimate: SE of slope (find using computer output)

$t^\star$: use a t-distribution with $df = n - 2$ and confidence level C%

(___, ___)

Conclude: Interpret the interval and, if applicable, draw a conclusion in context.

We are C% confident that the true *slope* of the regression line, the average change in [y] for each unit increase in [x], is between ___ and ___. If applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value 0.

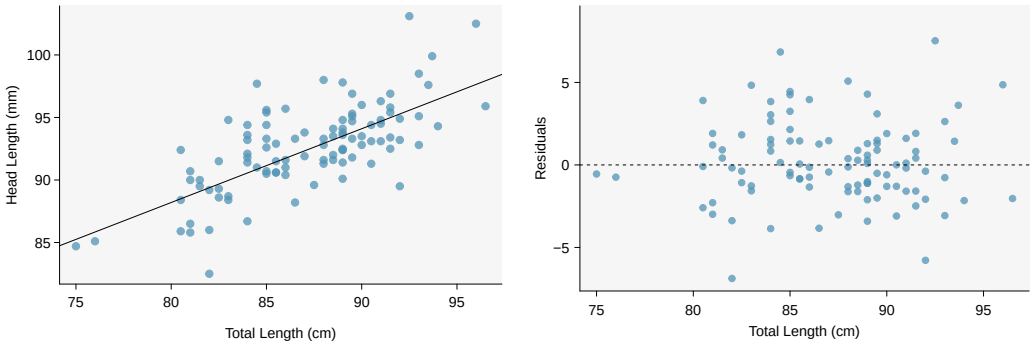


Figure 8.29: Left: Scatterplot of head length versus total length for 104 brushtail possums. Right: Residual plot for the model shown in left panel.

EXAMPLE 8.32

The regression summary below shows statistical software output from fitting the least squares regression line for predicting head length from total length for 104 brushtail possums. The scatterplot and residual plot are shown above.

```
Predictor        Coef        SE Coef    T        P
Constant         42.70979    5.17281    8.257    5.66e-13
total_length      0.57290    0.05933    9.657    4.68e-16

S = 2.595     R-Sq = 47.76%     R-Sq(adj) = 47.25%
```

Construct a 95% confidence interval for the slope of the regression line. Is there convincing evidence that there is a positive, linear relationship between head length and total length?

Identify: The parameter of interest is the slope of the population regression line for predicting head length from body length. We want to estimate this at the 95% confidence level.

Choose: Because the parameter to be estimated is the slope of a regression line, we will use the t-interval for the slope.

(E) **Check:** These data come from a random sample. The residual plot shows no pattern so a linear model seems reasonable. The residual plot also shows that the residuals have constant standard deviation. Finally, $n = 104 \geq 30$ so we do not have to check for skew in the residuals. All four conditions are met.

Calculate: We will calculate the interval: point estimate $\pm$ $t^\star \times SE$ of estimate

We read the slope of the sample regression line and the corresponding SE from the table. The point estimate is $b = 0.57290$. The SE of the slope is 0.05933, which can be found next to the slope of 0.57290. The degrees of freedom is $df = n - 2 = 104 - 2 = 102$. As before, we find the critical value $t^\star$ using a t-table (the $t^\star$ value is not the same as the T-statistic for the hypothesis test). Using the t-table at row $df = 100$ (round down since 102 is not on the table) and confidence level 95%, we get $t^\star = 1.984$.

So the 95% confidence interval is given by:

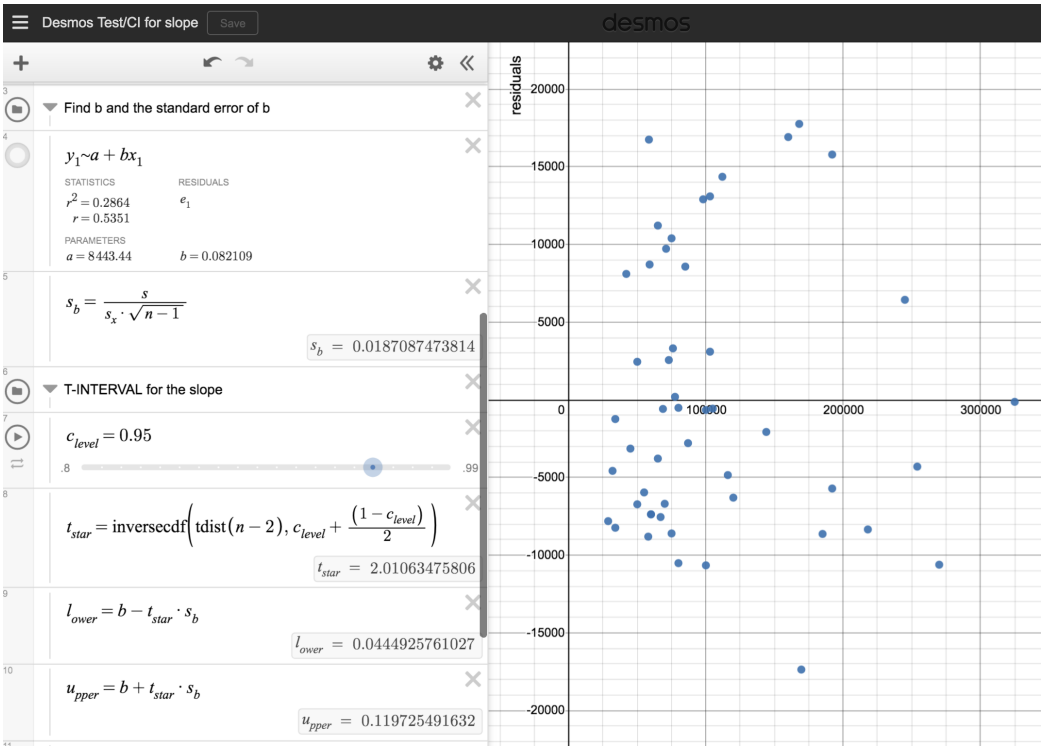
$$0.57290 \pm 1.984 \times 0.05933$$
$$(0.456, 0.691)$$

Conclude: We are 95% confident that the slope of the population regression line is between 0.456 and 0.691. That is, we are 95% confident that the true average *increase* in head length for each additional cm in total length is between 0.456 mm and 0.691 mm. Because the interval is entirely above 0, we do have evidence of a positive linear association between the head length and body length for brushtail possums.

8.4.4 Technology: the linear regression t-interval for the slope

We usually rely on regression output from statistical software when constructing confidence intervals for the slope of a regression line. However, it is also possible to use Desmos or a handheld calculator.

Get started quickly with this Desmos LinReg Calculator.



Entering a lot of data into a handheld calculator is not practical; we include TI instructions here simply for completion.

TI-84: T-INTERVAL FOR β

Use STAT, TESTS, LinRegTInt.

1. Choose STAT.
2. Right arrow to TESTS.
3. Down arrow and choose G: LinRegTInt.
   - This test is not built into the TI-83.
4. Let Xlist be L1 and Ylist be L2. (Don't forget to enter the x and y values in L1 and L2 before doing this interval.)
5. Let Freq be 1.
6. Enter the desired confidence level.
7. Leave RegEQ blank.
8. Choose Calculate and hit ENTER, which returns:

(__,__)	the confidence interval
b	b, the slope of best fit line of the sample data
df	degrees of freedom associated with this confidence interval
s	standard deviation of the residuals (not the same as SE of the slope)
a	a, the y-intercept of the best fit line of the sample data
r^2	R^2, the explained variance
r	r, the correlation coefficient

8.4.5 Midterm elections and unemployment

Elections for members of the United States House of Representatives occur every two years, coinciding every four years with the U.S. Presidential election. The set of House elections occurring during the middle of a Presidential term are called midterm elections. In America's two-party system, one political theory suggests the higher the unemployment rate, the worse the President's party will do in the midterm elections.

To assess the validity of this claim, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2018, with the exception of those elections during the Great Depression. Figure 8.30 shows these data and the least-squares regression line:

$$\begin{aligned}&\%\text{ change in House seats for President's party}\\&\quad = -7.36 - 0.89 \times (\text{unemployment rate})\end{aligned}$$

We consider the percent change in the number of seats of the President's party (e.g. percent change in the number of seats for Republicans in 2018) against the unemployment rate.

Examining the data, there are no clear deviations from linearity, the constant variance condition, or the normality of residuals. While the data are collected sequentially, a separate analysis was used to check for any apparent correlation between successive observations; no such correlation was found.

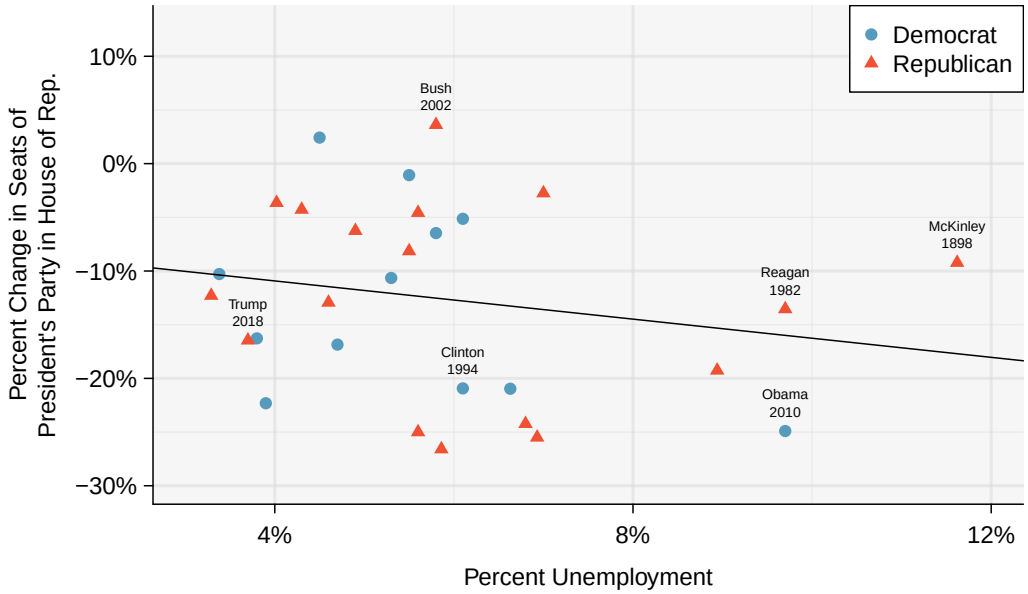


Figure 8.30: The percent change in House seats for the President's party in each election from 1898 to 2018 plotted against the unemployment rate. The two points for the Great Depression have been removed, and a least squares regression line has been fit to the data. Explore this data set on Tableau Public.

GUIDED PRACTICE 8.33

(G) The data for the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively. Do you agree that they should be removed for this investigation? Why or why not?[25]

There is a negative slope in the line shown in Figure 8.30. However, this slope (and the y-intercept) are only estimates of the parameter values. We might wonder, is this convincing evidence that the "true" linear model has a negative slope? That is, do the data provide strong evidence that the political theory is accurate? We can frame this investigation as a statistical hypothesis test:

H_0: $\beta = 0$. The true linear model has slope zero.

H_A: $\beta < 0$. The true linear model has a slope less than zero. The higher the unemployment, the greater the loss for the President's party in the House of Representatives.

We would reject H_0 in favor of H_A if the data provide strong evidence that the slope of the population regression line is less than zero. To assess the hypotheses, we identify a standard error for the estimate, compute an appropriate test statistic, and identify the p-value. Before we calculate these quantities, how good are we at visually determining from a scatterplot when a slope is significantly less than or greater than 0? And why do we tend to use a 0.05 significance level as our cutoff? Try out the following activity which will help answer these questions.

TESTING FOR THE SLOPE USING A CUTOFF OF 0.05

What does it mean to say that the slope of the population regression line is significantly greater than 0? And why do we tend to use a cutoff of $\alpha = 0.05$? This 5-minute interactive task will explain:

www.openintro.org/why05

[25] We will provide two considerations. Each of these points would have very high leverage on any least-squares regression line, and years with such high unemployment may not help us understand what would happen in other years where the unemployment is only modestly high. On the other hand, these are exceptional cases, and we would be discarding important information if we exclude them from a final analysis.

8.4.6 Understanding regression output from software

The residual plot shown in Figure 8.31 shows no pattern that would indicate that a linear model is inappropriate. Therefore we can carry out a test on the population slope using the sample slope as our point estimate. Just as for other point estimates we have seen before, we can compute a standard error and test statistic for b. The test statistic T follows a t-distribution with $n-2$ degrees of freedom.

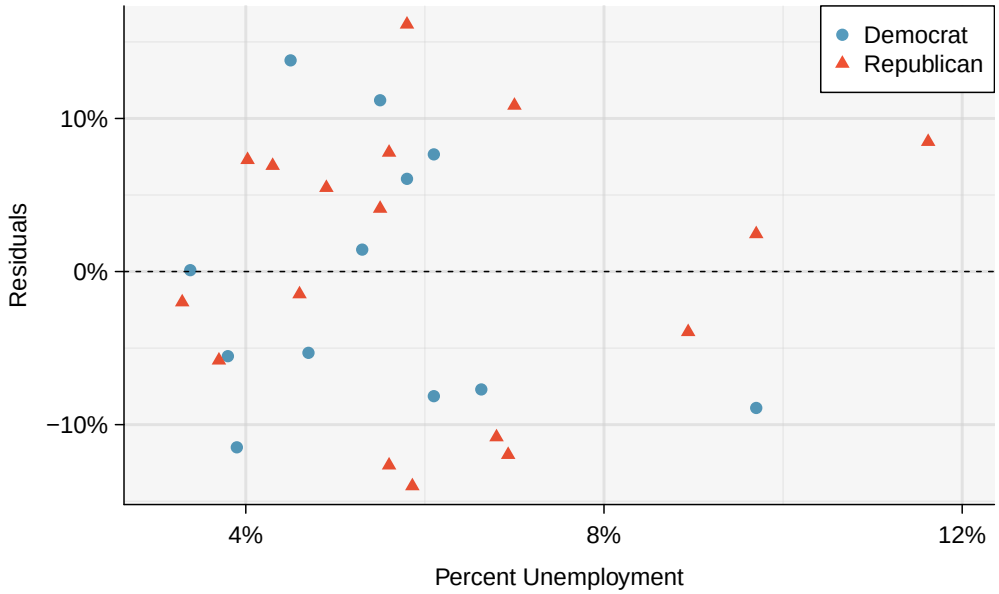


Figure 8.31: The residual plot shows no pattern that would indicate that a linear model is inappropriate. Explore this data set on Tableau Public.

HYPOTHESIS TESTS ON THE SLOPE OF THE REGRESSION LINE

Use a t-test with $n-2$ degrees of freedom when performing a hypothesis test on the slope of a regression line.

We will rely on statistical software to compute the standard error and leave the explanation of how this standard error is determined to a second or third statistics course. Figure 8.32 shows software output for the least squares regression line in Figure 8.30. The row labeled *unemp* represents the information for the slope, which is the coefficient of the unemployment variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.3644	5.1553	-1.43	0.1646
unemp	-0.8897	0.8350	-1.07	0.2961

Figure 8.32: Least squares regression summary for the percent change in seats of president's party in House of Reprepsentatives based on percent unemployment.

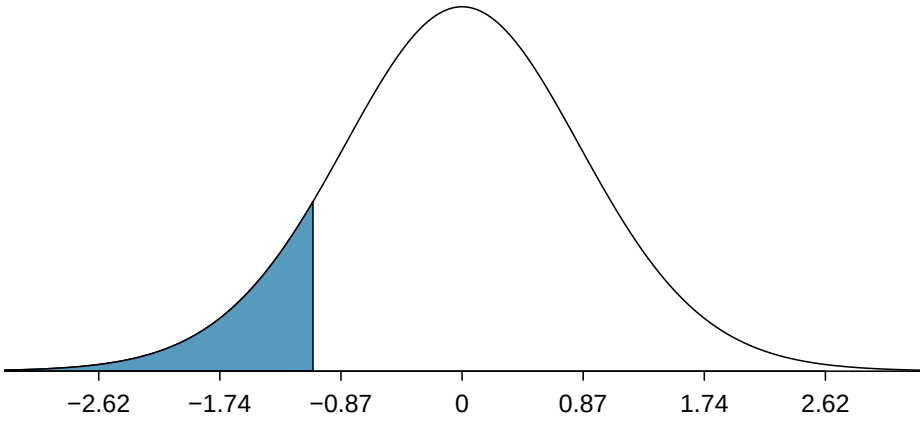


Figure 8.33: The distribution shown here is the sampling distribution for b, if the null hypothesis was true. The shaded tail represents the p-value for the hypothesis test evaluating whether there is convincing evidence that higher unemployment corresponds to a greater loss of House seats for the President's party during a midterm election.

EXAMPLE 8.34

What do the first column of numbers in the regression summary represent?

E

The entries in the first column represent the least squares estimates for the y-intercept and slope, a and b respectively. Using this information, we could write the equation for the least squares regression line as

$$\hat{y} = -7.3644 - 0.8897x$$

where y in this case represents the percent change in the number of seats for the president's party, and x represents the unemployment rate.

We previously used a test statistic T for hypothesis testing in the context of means. Regression is very similar. Here, the point estimate is $b = -0.8897$. The SE of the estimate is 0.8350, which is given in the second column, next to the estimate of b. This SE represents the typical error when using the slope of the sample regression line to estimate the slope of the population regression line.

The null value for the slope is 0, so we now have everything we need to compute the test statistic. We have:

$$T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}} = \frac{-0.8897 - 0}{0.8350} = -1.07$$

This value corresponds to the T-score reported in the regression output in the third column along the *unemp* row.

EXAMPLE 8.35

In this example, the sample size $n = 27$. Identify the degrees of freedom and p-value for the hypothesis test.

E

The degrees of freedom for this test is $n - 2$, or $df = 27 - 2 = 25$. We could use a table or a calculator to find the probability of a value less than -1.07 under the t-distribution with 25 degrees of freedom. However, the two-side p-value is given in Figure 8.32, next to the corresponding t-statistic. Because we have a one-sided alternative hypothesis, we take half of this. The p-value for the test is $\frac{0.2961}{2} = 0.148$.

Because the p-value is so large, we do not reject the null hypothesis. That is, the data do not provide convincing evidence that a higher unemployment rate is associated with a larger loss for the President's party in the House of Representatives in midterm elections.

DON'T CARELESSLY USE THE P-VALUE FROM REGRESSION OUTPUT

The last column in regression output often lists p-values for one particular hypothesis: a two-sided test where the null value is zero. If your test is one-sided and the point estimate is in the direction of H_A, then you can halve the software's p-value to get the one-tail area. If neither of these scenarios match your hypothesis test, be cautious about using the software output to obtain the p-value.

HYPOTHESIS TEST FOR THE SLOPE OF REGRESSION LINE

To carry out a complete hypothesis test for the claim that there is no linear relationship between two numerical variables, i.e. that $\beta = 0$,

Identify: Identify the hypotheses and the significance level, α.

H_0: $\beta = 0$

H_A: $\beta \neq 0$; H_A: $\beta > 0$; or H_A: $\beta < 0$

Choose: Choose the correct test procedure and identify it by name.

To test hypotheses about the slope of a regression model we use a ***t*-test for the slope**.

Check: Check conditions for using a t-test for the slope.

1. Independence: Data should come from a random sample or randomized experiment. If sampling without replacement, check that the sample size is less than 10% of the population size.
2. Linearity: Check that the scatterplot does not show a curved trend and that the residual plot shows no $\cup$-shape pattern.
3. Constant variability: Use the residual plot to check that the standard deviation of the residuals is constant across all x-values.
4. Normality: The population of residuals is nearly normal or the sample size is ≥ 30. If the sample size is less than 30 check for strong skew or outliers in the sample residuals. If neither is found, then the condition that the population of residuals is nearly normal is considered reasonable.

Calculate: Calculate the t-statistic, df, and p-value.

$$T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}, \quad df = n - 2$$

point estimate: the slope b of the sample regression line

SE of estimate: SE of slope (find using computer output)

null value: 0

p-value = (based on the t-statistic, the df, and the direction of H_A)

Conclude: Compare the p-value to α, and draw a conclusion in context.

If the p-value is $< \alpha$, reject H_0; there is sufficient evidence that [H_A in context].

If the p-value is $> \alpha$, do not reject H_0; there is not sufficient evidence that [H_A in context].

EXAMPLE 8.36

The regression summary below shows statistical software output from fitting the least squares regression line for predicting gift aid based on family income for 50 randomly selected freshman students at Elmhurst College. The scatterplot and residual plot were shown in Figure 8.27.

```
Predictor         Coef         SE Coef    T         P
Constant          24.31933     1.29145    18.831    < 2e-16
family_income     -0.04307     0.01081    -3.985    0.000229

S = 4.783     R-Sq = 24.86%     R-Sq(adj) = 23.29%
```

Do these data provide convincing evidence that there is a negative, linear relationship between family income and gift aid? Carry out a complete hypothesis test at the 0.05 significance level. Use the five step framework to organize your work.

Identify: We will test the following hypotheses at the $\alpha = 0.05$ significance level.
H_0: $\beta = 0$. There is no linear relationship.
H_A: $\beta < 0$. There is a negative linear relationship.

Here, β is the slope of the population regression line for predicting gift aid from family income at Elmhurst College.

(E) **Choose:** Because the hypotheses are about the slope of a regression line, we choose the t-test for a slope.

Check: The data come from a random sample of less than 10% of the total population of freshman students at Elmhurst College. The lack of any pattern in the residual plot indicates that a linear model is reasonable. Also, the residual plot shows that the residuals have constant variance. Finally, $n = 50 \geq 30$ so we do not have to worry too much about any skew in the residuals. All four conditions are met.

Calculate: We will calculate the t-statistic, degrees of freedom, and the p-value.

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

We read the slope of the sample regression line and the corresponding SE from the table.
The point estimate is: $b = -0.04307$.
The SE of the slope is: $SE = 0.01081$.

$$T = \frac{-0.04307 - 0}{0.01081} = -3.985$$

Because H_A uses a less than sign ($<$), meaning that it is a lower-tail test, the p-value is the area to the *left* of $t = -3.985$ under the t-distribution with $50 - 2 = 48$ degrees of freedom. The p-value $= \frac{1}{2}(0.000229) \approx 0.0001$.

Conclude: The p-value of 0.0001 is < 0.05, so we reject H_0; there is sufficient evidence that there is a negative linear relationship between family income and gift aid at Elmhurst College.

(G) **Guided Practice 8.37**

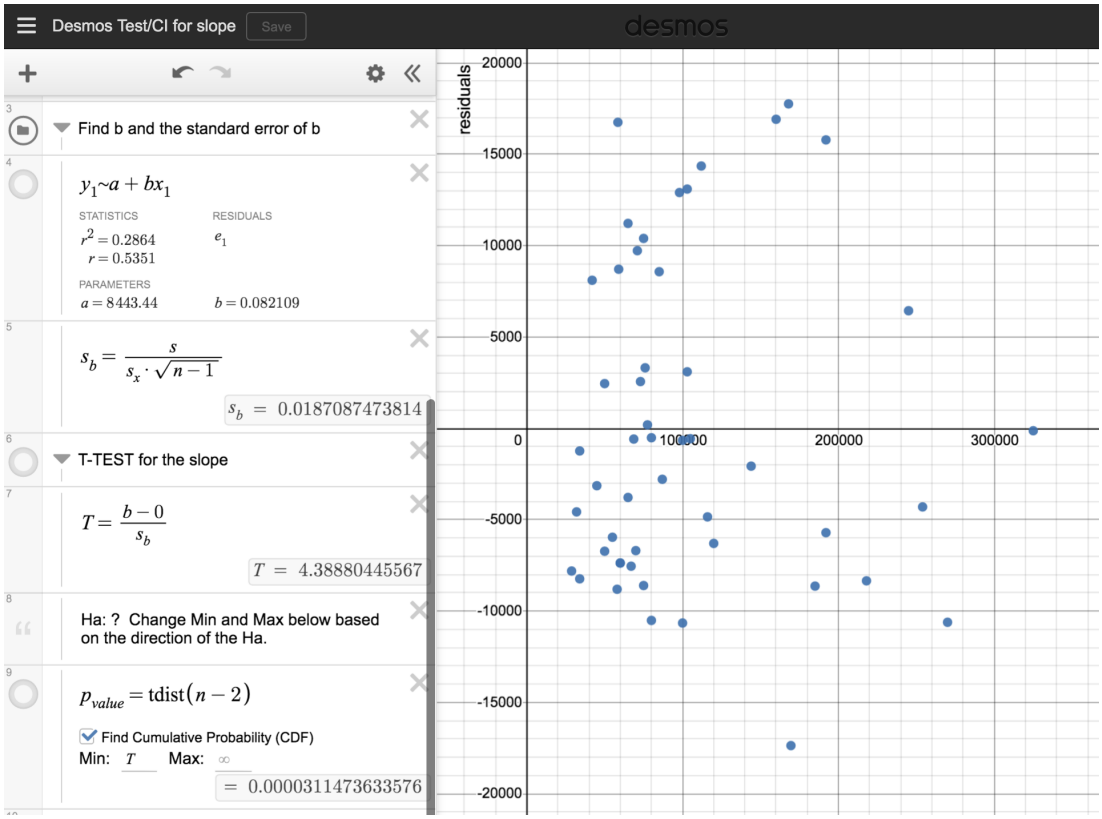
In context, interpret the p-value from the previous example.[26]

[26] Assuming that the probability model is true and assuming that the null hypothesis is true, i.e. there really is no linear relationship between family income and gift aid at Elmhurst College, there is only a 0.0001 chance of getting a test statistic this small or smaller (H_A uses a $<$, so the p-value represents the area in the left tail). Because this value is so small, we reject the null hypothesis.

8.4.7 Technology: the t-test for the slope

We generally rely on regression output from statistical software programs to provide us with the necessary quantities: b and SE of b. However we can also find the test statistic and p-value using Desmos or a handheld calculator.

Get started quickly with this Desmos T-Test/Interval Calculator.



Entering a lot of data into a handheld calculator is not practical; we include TI instructions here simply for completion.

TI-83/84: LINEAR REGRESSION T-TEST FOR β

Use STAT, TESTS, LinRegTTest.

1. Choose STAT.
2. Right arrow to TESTS.
3. Down arrow and choose F:LinRegTTest. (On TI-83 it is E:LinRegTTest).
4. Let Xlist be L1 and Ylist be L2. (Don't forget to enter the x and y values in L1 and L2 before doing this test.)
5. Let Freq be 1.
6. Choose $\neq$, $<$, or $>$ to correspond to H_A.
7. Leave RegEQ blank.
8. Choose Calculate and hit ENTER, which returns:

t	t statistic	b	b, slope of the line
p	p-value	s	st. dev. of the residuals
df	degrees of freedom for the test	r^2	R^2, explained variance
a	a, y-intercept of the line	r	r, correlation coefficient

CASIO FX-9750GII: LINEAR REGRESSION T-TEST FOR β

1. Navigate to `STAT` (`MENU` button, then hit the `2` button or select `STAT`).
2. Enter your data into 2 lists.
3. Select `TEST` (`F3`), `t` (`F2`), and `REG` (`F3`).
4. If needed, update the sidedness of the test and the `XList` and `YList` lists. The `Freq` should be set to `1`.
5. Hit `EXE`, which returns:

`t`	t statistic	`b`	b, slope of the line
`p`	p-value	`s`	st. dev. of the residuals
`df`	degrees of freedom for the test	`r`	r, correlation coefficient
`a`	a, y-intercept of the line	`r²`	R^2, explained variance

EXAMPLE 8.38

Why does the calculator test include the symbol ρ when choosing the direction of the alternative hypothesis?

(E)

Recall the we used the letter r to represent correlation. The Greek letter $\rho = 0$ represents the correlation for the entire population. The slope $b = r\frac{s_y}{s_x}$. If the slope of the population regression line is zero, the correlation for the population must also be zero. For this reason, the t-test for $\beta = 0$ is equivalent to a test for $\rho = 0$.

8.4.8 Which inference procedure to use for paired data?

In Section **??**, we looked at a set of paired data involving the price of textbooks for UCLA courses at the UCLA Bookstore and on Amazon. The left panel of Figure 8.34 shows the difference in price (UCLA Bookstore − Amazon) for each book. Because we have two data points on each textbook, it also makes sense to construct a scatterplot, as seen in the right panel of Figure 8.34.

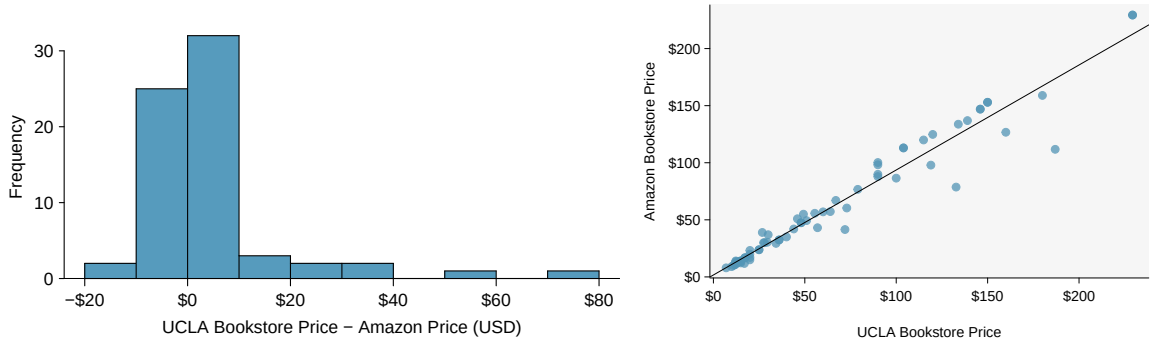


Figure 8.34: Left: histogram of the difference (UCLA Bookstore - Amazon) in price for each book sampled. Right: scatterplot of Amazon Price versus UCLA Bookstore price.

EXAMPLE 8.39

What additional information does the scatterplot provide about the price of textbooks at UCLA Bookstore and on Amazon?

(E)

With a scatterplot, we see the *relationship* between the variables. We can see when UCLA Bookstore price is larger, whether Amazon price tends to be larger. We can consider the strength of the correlation and we can plot the linear regression equation.

EXAMPLE 8.40

Which test should we do if we want to check whether:

1. prices for textbooks for UCLA courses are *higher* at the UCLA Bookstore than on Amazon
2. there is a significant, positive linear relationship between UCLA Bookstore price and Amazon price?

(E)

In the first case, we are interested in whether the differences (UCLA Bookstore − Amazon) are, on average, greater than 0, so we would do a paired t-test for a mean of differences. In the second case, we are interested in whether the slope is significantly greater than 0, so we would do a t-test for the slope of a regression line.

Likewise, a paired t-interval for a mean of differences would provide an interval of reasonable values for mean of the differences for all UCLA textbooks, whereas a t-interval for the slope would provide an interval of reasonable values for the slope of the regression line for all UCLA textbooks.

INFERENCE FOR PAIRED DATA

A paired t-interval or t-test for a mean of differences only makes sense when we are asking whether, on average, one variable is *greater* than another (think histogram of the differences). A t-interval or t-test for the slope of a regression line makes sense when we are interested in the linear relationship between them (think scatterplot).

EXAMPLE 8.41

Previously, we looked at the relationship betweeen body length and head length for bushtail possums. We also looked at the relationship between gift aid and family income for freshmen at Elmhurst College. Could we do a paired t-test in either of these scenarios?

(E)

We have to ask ourselves, does it make sense to ask whether, on average, body length is greater than head length? Similarly, does it make sense to ask whether, on average, gift aid is greater than family income? These don't seem to be meaningful research questions; a paired t-test for a mean of differences would not be useful here.

GUIDED PRACTICE 8.42

(G) A teacher gives her class a pretest and a posttest. Does this result in paired data? If so, which hypothesis test should she use?[27]

[27] Yes, there are two observations for each individual, so there is paired data. The appropriate test depends upon the question she wants to ask. If she is interested in whether, on average, students do better on the posttest than the pretest, should use a paired t-test for a mean of differences. If she is interested in whether pretest score is a significant linear predictor of posttest score, she should do a t-test for the slope. In this situation, both tests could be useful, but which one should be used is dependent on the teacher's research question.

Section summary

In Chapter 6, we used a χ^2 test of independence to test for association between two categorical variables. In this section, we test for association/correlation between two numerical variables.

- We use the slope b as a *point estimate* for the slope β of the population regression line. The slope of the population regression line is the true increase/decrease in y for each unit increase in x. If the slope of the population regression line is 0, there is no linear relationship between the two variables.
- Under certain assumptions, the sampling distribution of b is *normal* and the distribution of the standardized test statistic using the standard error of the slope follows a ***t*-distribution** with $n-2$ degrees of freedom.
- When there is (x, y) data and the parameter of interest is the slope of the population regression line, e.g. the slope of the population regression line relating air quality index to average rainfall per year for each city in the United States:
  - Estimate β at the C% confidence level using a ***t*-interval for the slope**.
  - Test H_0: $\beta = 0$ at the α significance level using a ***t*-test for the slope**.
- The conditions for the t-interval and t-test for the slope of a regression line are the same.
  1. Independence: Data come from a random sample or randomized experiment. If sampling without replacement, check that the sample size is less than 10% of the population size.
  2. Linearity: Check that the scatterplot does not show a curved trend and that the residual plot shows no $\cup$-shape pattern.
  3. Constant variability: Use the residual plot to check that the standard deviation of the residuals is constant across all x-values.
  3. Normality: The population of residuals is nearly normal or the sample size is ≥ 30. If the sample size is less than 30 check for strong skew or outliers in the sample residuals. If neither is found, then the condition that the population of residuals is nearly normal is considered reasonable.
- The confidence interval and test statistic are calculated as follows:

  Confidence interval: point estimate $\pm$ $t^\star \times SE$ of estimate, or

  Test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$ and p-value

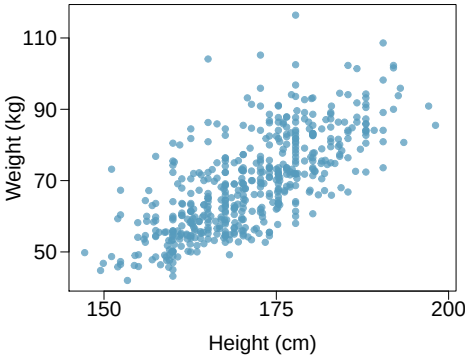
  point estimate: the slope b of the sample regression line

  SE of estimate: SE of slope (find using computer output)

  $df = n - 2$
- If the confidence interval for the slope of the population regression line estimates the true average increase in the y-variable for each unit increase in the x-variable.
- The t-test for the slope and the paired t-test for a mean of differences both involve *paired*, numerical data. However, the t-test for the slope asks if the two variables have a linear *relationship*, specifically if the *slope* of the population regression line is different from 0. The paired t-test for a mean of differences, on the other hand, asks if the two variables are in some way the *same*, specifically if the *mean* of the population differences is 0.

Exercises

8.33 Body measurements, Part IV. The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-105.0113	7.5394	-13.93	0.0000
height	1.0176	0.0440	23.13	0.0000

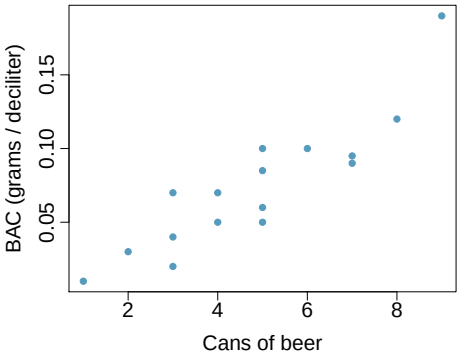
(a) Describe the relationship between height and weight.

(b) Write the equation of the regression line. Interpret the slope and intercept in context.

(c) Do the data provide strong evidence that an increase in height is associated with an increase in weight? State the null and alternative hypotheses, report the p-value, and state your conclusion.

(d) The correlation coefficient for height and weight is 0.72. Calculate R^2 and interpret it in context.

8.34 Beer and blood alcohol content. Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were different genders, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood.[28] The scatterplot and regression table summarize the findings.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0127	0.0126	-1.00	0.3320
beers	0.0180	0.0024	7.48	0.0000

(a) Describe the relationship between the number of cans of beer and BAC.

(b) Write the equation of the regression line. Interpret the slope and intercept in context.

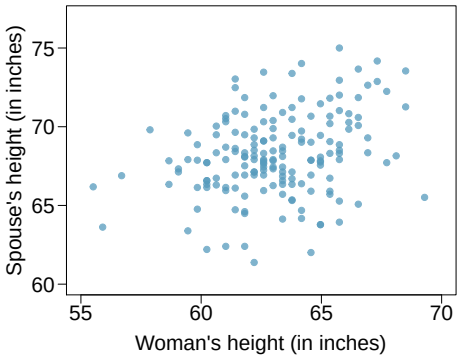
(c) Do the data provide strong evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion.

(d) The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R^2 and interpret it in context.

(e) Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study?

[28] **Malkevitc+Lesser:2008**.

8.35 Spouses, Part II. The scatterplot below summarizes womens' heights and their spouses' heights for a random sample of 170 married women in Britain, where both partners' ages are below 65 years. Summary output of the least squares fit for predicting spouse's height from the woman's height is also provided in the table.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	43.5755	4.6842	9.30	0.0000
height_spouse	0.2863	0.0686	4.17	0.0000

(a) Is there strong evidence in this sample that taller women have taller spouses? State the hypotheses and include any information used to conduct the test.

(b) Write the equation of the regression line for predicting the height of a woman's spouse based on the woman's height.

(c) Interpret the slope and intercept in the context of the application.

(d) Given that $R^2 = 0.09$, what is the correlation of heights in this data set?

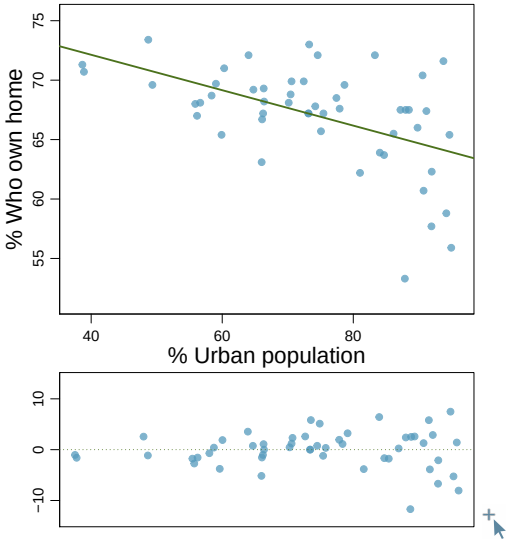
(e) You meet a married woman from Britain who is 5'9" (69 inches). What would you predict her spouse's height to be? How reliable is this prediction?

(f) You meet another married woman from Britain who is 6'7" (79 inches). Would it be wise to use the same linear model to predict her spouse's height? Why or why not?

8.36 Urban homeowners, Part II. Exercise 8.29 gives a scatterplot displaying the relationship between the percent of families that own their home and the percent of the population living in urban areas. Below is a similar scatterplot, excluding District of Columbia, as well as the residuals plot. There were 51 cases.

(a) For these data, $R^2 = 0.28$. What is the correlation? How can you tell if it is positive or negative?

(b) Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for these data?



8.37 Murders and poverty, Part II. Exercise 8.25 presents regression output from a model for predicting annual murders per million from percentage living in poverty based on a random sample of 20 metropolitan areas. The model output is also provided below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$$s = 5.512 \qquad R^2 = 70.52\% \qquad R^2_{adj} = 68.89\%$$

(a) What are the hypotheses for evaluating whether poverty percentage is a significant predictor of murder rate?

(b) State the conclusion of the hypothesis test from part (a) in context of the data.

(c) Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context of the data.

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

8.38 Babies. Is the gestational age (time between conception and birth) of a low birth-weight baby useful in predicting head circumference at birth? Twenty-five low birth-weight babies were studied at a Harvard teaching hospital; the investigators calculated the regression of head circumference (measured in centimeters) against gestational age (measured in weeks). The estimated regression line is

$$\widehat{head\ circumference} = 3.91 + 0.78 \times gestational\ age$$

(a) What is the predicted head circumference for a baby whose gestational age is 28 weeks?

(b) The standard error for the coefficient of gestational age is 0. 35, which is associated with $df = 23$. Does the model provide strong evidence that gestational age is significantly associated with head circumference?

Chapter highlights

This chapter focused on describing the linear association between two numerical variables and fitting a linear model.

- The **correlation coefficient**, r, measures the strength and direction of the linear association between two variables. However, r alone cannot tell us whether data follow a linear trend or whether a linear model is appropriate.
- The **explained variance**, R^2, measures the proportion of variation in the y values explained by a given model. Like r, R^2 alone cannot tell us whether data follow a linear trend or whether a linear model is appropriate.
- Every analysis should begin with *graphing* the data using a **scatterplot** in order to see the association and any deviations from the trend (outliers or influential values). A **residual plot** helps us better see patterns in the data.
- When the data show a linear trend, we fit a **least squares regression line** of the form: $\hat{y} = a + bx$, where a is the y-intercept and b is the slope. It is important to be able to *calculate* a and b using the summary statistics and to *interpret* them in the context of the data.
- A **residual**, $y - \hat{y}$, measures the error for an *individual point*. The **standard deviation of the residuals**, s, measures the typical size of the residuals.
- $\hat{y} = a + bx$ provides the best fit line for the *observed data*. To estimate or hypothesize about the slope of the population regression line, first confirm that the residual plot has no pattern and that a linear model is reasonable, then use a ***t*-interval for the slope** or a ***t*-test for the slope** with $n - 2$ degrees of freedom.

In this chapter we focused on simple linear models with one explanatory variable. More complex methods of prediction, such as multiple regression (more than one explanatory variable) and non-linear regression can be studied in a future course.

Chapter exercises

8.39 True / False. Determine if the following statements are true or false. If false, explain why.

(a) A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation of 0.5.

(b) Correlation is a measure of the association between any two variables.

8.40 Cats, Part II. Exercise 8.26 presents regression output from a model for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cat. The model output is also provided below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000

$$s = 1.452 \qquad R^2 = 64.66\% \qquad R^2_{adj} = 64.41\%$$

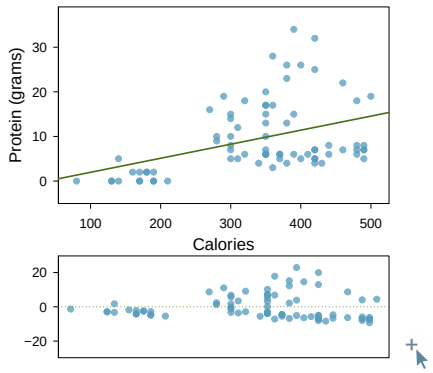
(a) We see that the point estimate for the slope is positive. What are the hypotheses for evaluating whether body weight is positively associated with heart weight in cats?

(b) State the conclusion of the hypothesis test from part (a) in context of the data.

(c) Calculate a 95% confidence interval for the slope of body weight, and interpret it in context of the data.

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

8.41 Nutrition at Starbucks, Part II. Exercise 8.22 introduced a data set on nutrition information on Starbucks food menu items. Based on the scatterplot and the residual plot provided, describe the relationship between the protein content and calories of these menu items, and determine if a simple linear model is appropriate to predict amount of protein from the number of calories.



8.42 Helmets and lunches. The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (`lunch`) and the percentage of bike riders in the neighborhood wearing helmets (`helmet`). The average percentage of children receiving reduced-fee lunches is 30.8% with a standard deviation of 26.7% and the average percentage of bike riders wearing helmets is 38.8% with a standard deviation of 16.9%.

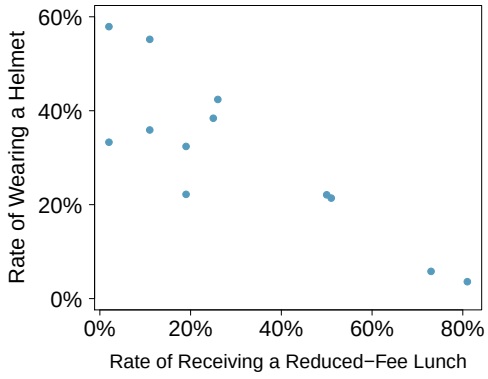
(a) If the R^2 for the least-squares regression line for these data is 72%, what is the correlation between `lunch` and `helmet`?

(b) Calculate the slope and intercept for the least-squares regression line for these data.

(c) Interpret the intercept of the least-squares regression line in the context of the application.

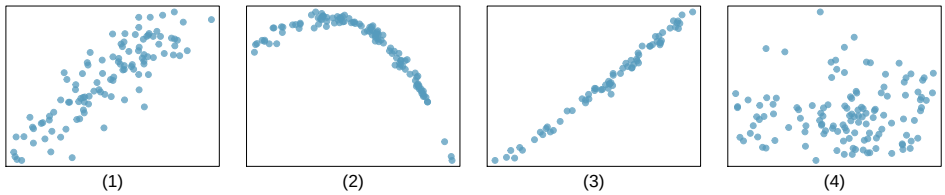
(d) Interpret the slope of the least-squares regression line in the context of the application.

(e) What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual in the context of the application.



8.43 Match the correlation, Part III. Match each correlation to the corresponding scatterplot.

(a) $r = -0.72$
(b) $r = 0.07$
(c) $r = 0.86$
(d) $r = 0.99$



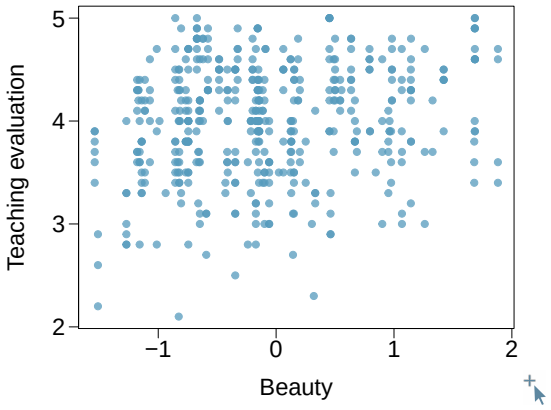
8.44 Rate my professor. Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors.[29] The scatterplot below shows the relationship between these variables, and regression output is provided for predicting teaching evaluation score from beauty score.

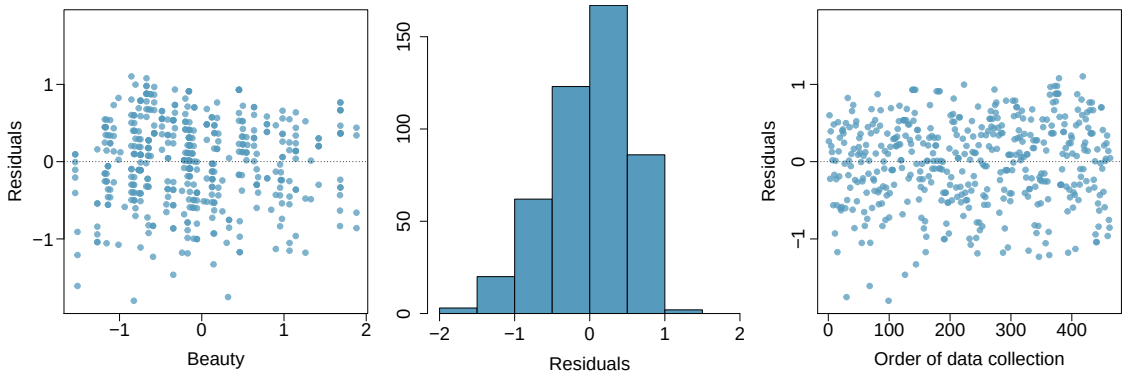
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty		0.0322	4.13	0.0000

(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

(b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

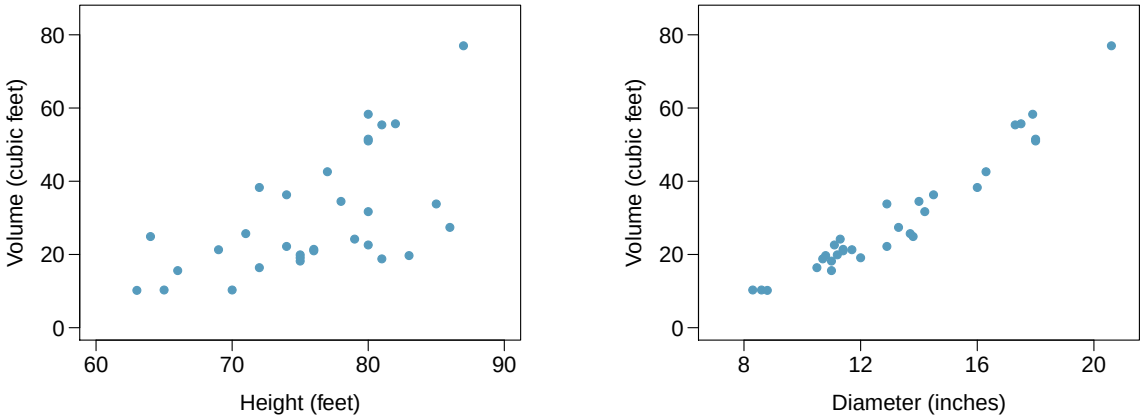
(c) List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.





[29] **Hamermesh:2005**.

8.45 Trees. The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground.[30]



(a) Describe the relationship between volume and height of these trees.

(b) Describe the relationship between volume and diameter of these trees.

(c) Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

[30] **data:trees**.

Appendix A

Exercise solutions

1 Data collection

?? (a) Treatment: $10/43 = 0.23 \rightarrow 23\%$.
(b) Control: $2/46 = 0.04 \rightarrow 4\%$. (c) A higher percentage of patients in the treatment group were pain free 24 hours after receiving acupuncture. (d) It is possible that the observed difference between the two group percentages is due to chance.

?? (a) "Is there an association between air pollution exposure and preterm births?" (b) 143,196 births in Southern California between 1989 and 1993. (c) Measurements of carbon monoxide, nitrogen dioxide, ozone, and particulate matter less than $10\mu g/m^3$ (PM_{10}) collected at air-quality-monitoring stations as well as length of gestation. Continuous numerical variables.

?? (a) "Does explicitly telling children not to cheat affect their likelihood to cheat?". (b) 160 children between the ages of 5 and 15. (c) Four variables: (1) age (numerical, continuous), (2) sex (categorical), (3) whether they were an only child or not (categorical), (4) whether they cheated or not (categorical).

?? Explanatory: acupuncture or not. Response: if the patient was pain free or not.

?? (a) $50 \times 3 = 150$. (b) Four continuous numerical variables: sepal length, sepal width, petal length, and petal width. (c) One categorical variable, species, with three levels: *setosa*, *versicolor*, and *virginica*.

?? (a) Airport ownership status (public/private), airport usage status (public/private), latitude, and longitude. (b) Airport ownership status: categorical, not ordinal. Airport usage status: categorical, not ordinal. Latitude: numerical, continuous. Longitude: numerical, continuous.

?? (a) Population: all births, sample: 143,196 births between 1989 and 1993 in Southern California. (b) If births in this time span at the geography can be considered to be representative of all births, then the results are generalizable to the population of Southern California. However, since the study is observational the findings cannot be used to establish causal relationships.

?? (a) Population: all asthma patients aged 18-69 who rely on medication for asthma treatment. Sample: 600 such patients. (b) If the patients in this sample, who are likely not randomly sampled, can be considered to be representative of all asthma patients aged 18-69 who rely on medication for asthma treatment, then the results are generalizable to the population defined above. Additionally, since the study is experimental, the findings can be used to establish causal relationships.

?? (a) Observation. (b) Variable. (c) Sample statistic (mean). (d) Population parameter (mean).

?? (a) Observational. (b) Use stratified sampling to randomly sample a fixed number of students, say 10, from each section for a total sample size of 40 students.

?? (a) Positive, non-linear, somewhat strong. Countries in which a higher percentage of the population have access to the internet also tend to have higher average life expectancies, however rise in life expectancy trails off before around 80 years old. (b) Observational. (c) Wealth: countries with individuals who can widely afford the internet can probably also afford basic medical care. (Note: Answers may vary.)

?? (a) Simple random sampling is okay. In fact, it's rare for simple random sampling to not be a reasonable sampling method! (b) The student opinions may vary by field of study, so the stratifying by this variable makes sense and would be reasonable. (c) Students of similar ages are probably going to have more similar opinions, and we want clusters to be diverse with respect to the outcome of interest, so this would **not** be a good approach. (Additional thought: the clusters in this case may also have very different numbers of people, which can also create unexpected sample sizes.)

?? (a) The cases are 200 randomly sampled men and women. (b) The response variable is attitude towards a fictional microwave oven. (c) The explanatory variable is dispositional attitude. (d) Yes, the cases are sampled randomly. (e) This is an observational study since there is no random assignment to treatments. (f) No, we cannot establish a causal link between the explanatory and response variables since the study is observational. (g) Yes, the results of the study can be generalized to the population at large since the sample is random.

?? (a) Simple random sample. Non-response bias, if only those people who have strong opinions about the survey responds his sample may not be representative of the population. (b) Convenience sample. Under coverage bias, his sample may not be representative of the population since it consists only of his friends. It is also possible that the study will have non-response bias if some choose to not bring back the survey. (c) Convenience sample. This will have a similar issues to handing out surveys to friends. (d) Multi-stage sampling. If the classes are similar to each other with respect to student composition this approach should not introduce bias, other than potential non-response bias.

?? (a) Exam performance. (b) Light level: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). (c) Sex: man, woman.

?? (a) Exam performance. (b) Light level (overhead lighting, yellow overhead lighting, no overhead lighting) and noise level (no noise, construction noise, and human chatter noise). (c) Since the researchers want to ensure equal gender representation, sex will be a blocking variable.

?? Need randomization and blinding. One possible outline: (1) Prepare two cups for each participant, one containing regular Coke and the other containing Diet Coke. Make sure the cups are identical and contain equal amounts of soda. Label the cups A (regular) and B (diet). (Be sure to randomize A and B for each trial!) (2) Give each participant the two cups, one cup at a time, in random order, and ask the participant to record a value that indicates how much she liked the beverage. Be sure that neither the participant nor the person handing out the cups knows the identity of the beverage to make this a double- blind experiment. (Answers may vary.)

?? (a) Observational study. (b) Dog: Lucy. Cat: Luna. (c) Oliver and Lily. (d) Positive, as the popularity of a name for dogs increases, so does the popularity of that name for cats.

?? (a) Experiment. (b) Treatment: 25 grams of chia seeds twice a day, control: placebo. (c) Yes, gender. (d) Yes, single blind since the patients were blinded to the treatment they received. (e) Since this is an experiment, we can make a causal statement. However, since the sample is not random, the causal statement cannot be generalized to the population at large.

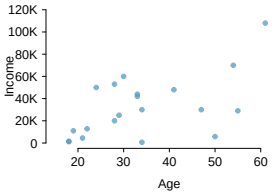
?? (a) Non-responders may have a different response to this question, e.g. parents who returned the surveys likely don't have difficulty spending time with their children. (b) It is unlikely that the women who were reached at the same address 3 years later are a random sample. These missing responders are probably renters (as opposed to homeowners) which means that they might be in a lower socio- economic status than the respondents. (c) There is no control group in this study, this is an observational study, and there may be confounding variables, e.g. these people may go running because they are generally healthier and/or do other exercises.

?? (a) Randomized controlled experiment. (b) Explanatory: treatment group (categorical, with 3 levels). Response variable: Psychological well-being. (c) No, because the participants were volunteers. (d) Yes, because it was an experiment. (e) The statement should say "evidence" instead of "proof".

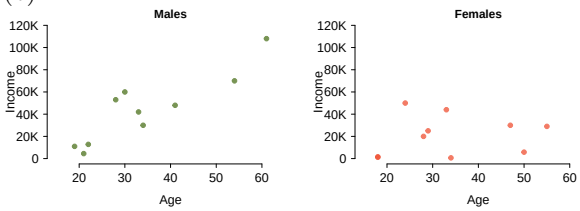
?? (a) County, state, driver's race, whether the car was searched or not, and whether the driver was arrested or not. (b) All categorical, non-ordinal. (c) Response: whether the car was searched or not. Explanatory: race of the driver.

2 Summarizing data

2.1 (a) There is a weak and positive relationship between age and income. With so few points it is difficult to tell the form of the relationship (linear or not) however the relationship does look somewhat curved.



(b)

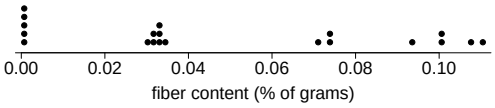


(c) For males as age increases so does income, however this pattern is not apparent for females.

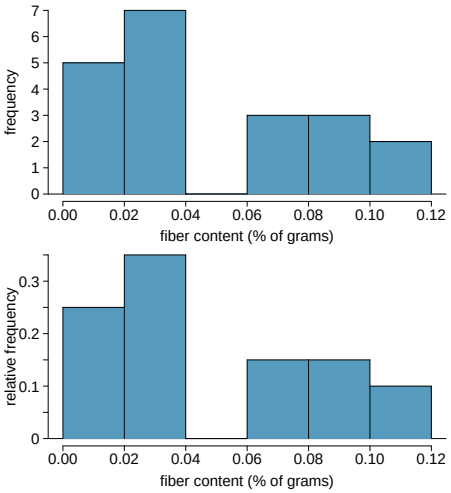
2.3 (a)

```
0 | 000003333333
0 | 7779
1 | 0011
Legend:  1 | 0 = 10%
```

(b)



(c)



(d) 40% (Note: if using only rel. freq. histogram, you can only get an estimate because 7 is in the middle of the bin. Use the dot plot to get a more accurate answer.)

2.5 (a) Positive association: mammals with longer gestation periods tend to live longer as well. (b) Association would still be positive. (c) No, they are not independent. See part (a).

2.7 Both distributions are right skewed and bimodal with modes at 10 and 20 cigarettes; note that people may be rounding their answers to half a pack or a whole pack. The median of each distribution is between 10 and 15 cigarettes. The middle 50% of the data (the IQR) appears to be spread equally in each group and have a width of about 10 to 15. There are potential outliers above 40 cigarettes per day. It appears that respondents who smoke only a few cigarettes (0 to 5) smoke more on the weekdays than on weekends.

2.9 (a) $\bar{x}_{amtWeekends} = 20$, $\bar{x}_{amtWeekdays} = 16$. (b) $s_{amtWeekends} = 0$, $s_{amtWeekdays} = 4.18$. In this very small sample, higher on weekdays.

2.11 Any 10 employees whose average number of days off is between the minimum and the mean number of days off for the entire workforce at this plant.

2.13 (a) Dist 2 has a higher mean since $20 > 13$, and a higher standard deviation since 20 is further from the rest of the data than 13. (b) Dist 1 has a higher mean since $-20 > -40$, and Dist 2 has a higher standard deviation since -40 is farther away from the rest of the data than -20. (c) Dist 2 has a higher mean since all values in this distribution are higher than those in Dist 1, but both distribution have the same standard deviation since they are equally variable around their respective means. (d) Both distributions have the same mean since they're both centered at 300, but Dist 2 has a higher standard deviation since the observations are farther from the mean than in Dist 1.

2.15 (a) About 30. (b) Since the distribution is right skewed the mean is higher than the median. (c) Q1: between 15 and 20, Q3: between 35 and 40, IQR: about 20. (d) Values that are considered to be unusually low or high lie more than 1.5×IQR away from the quartiles. Upper fence: Q3 + 1.5 × IQR = $37.5 + 1.5 \times 20 = 67.5$; Lower fence: Q1 - 1.5 × IQR = $17.5 + 1.5 \times 20 = -12.5$; The lowest AQI recorded is not lower than 5 and the highest AQI recorded is not higher than 65, which are both within the fences. Therefore none of the days in this sample would be considered to have an unusually low or high AQI.

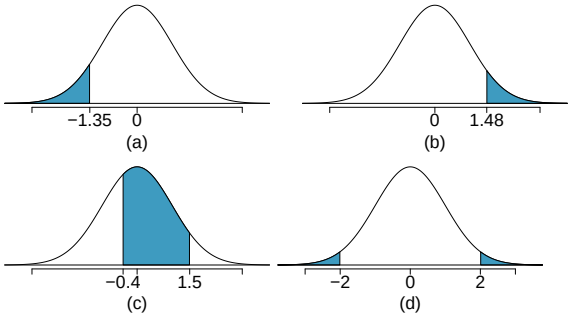
2.17 The histogram shows that the distribution is bimodal, which is not apparent in the box plot. The box plot makes it easy to identify more precise values of observations outside of the whiskers.

2.19 (a) The distribution of number of pets per household is likely right skewed as there is a natural boundary at 0 and only a few people have many pets. Therefore the center would be best described by the median, and variability would be best described by the IQR. (b) The distribution of number of distance to work is likely right skewed as there is a natural boundary at 0 and only a few people live a very long distance from work. Therefore the center would be best described by the median, and variability would be best described by the IQR. (c) The distribution of heights of males is likely symmetric. Therefore the center would be best described by the mean, and variability would be best described by the standard deviation.

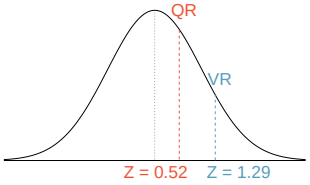
2.21 (a) The median is a much better measure of the typical amount earned by these 42 people. The mean is much higher than the income of 40 of the 42 people. This is because the mean is an arithmetic average and gets affected by the two extreme observations. The median does not get effected as much since it is robust to outliers. (b) The IQR is a much better measure of variability in the amounts earned by nearly all of the 42 people. The standard deviation gets affected greatly by the two high salaries, but the IQR is robust to these extreme observations.

2.23 (a) The distribution is unimodal and symmetric with a mean of about 25 minutes and a standard deviation of about 5 minutes. There does not appear to be any counties with unusually high or low mean travel times. (b) Answers will vary. There are pockets of longer travel time around DC, Southeastern NY, Chicago, Minneapolis, Los Angeles, and many other big cities. There is also a large section of shorter average commute times that overlap with farmland in the Midwest. Many farmers' homes are adjacent to their farmland, so their commute would be brief, which may explain why the average commute time for these counties is relatively low.

2.25 (a) 8.85%. (b) 6.94%. (c) 58.86%. (d) 4.56%.



2.27 (a) $Z_{VR} = 1.29$, $Z_{QR} = 0.52$.



(b) She scored 1.29 standard deviations above the mean on the Verbal Reasoning section and 0.52 standard deviations above the mean on the Quantitative Reasoning section. (c) She did better on the Verbal Reasoning section since her Z-score on that section was higher. (d) $Perc_{VR} = 0.9007 \approx 90\%$, $Perc_{QR} = 0.6990 \approx 70\%$. (e) $100\% - 90\% = 10\%$ did better than her on VR, and $100\% - 70\% = 30\%$ did better than her on QR. (f) We cannot compare the raw scores since they are on different scales. Comparing her percentile scores is more appropriate when comparing her performance to others. (g) Answer to part (b) would not change as Z-scores can be calculated for distributions that are not normal. However, we could not answer parts (c)-(e) since we cannot use the normal probability table to calculate probabilities and percentiles without a normal model.

2.29 (a) $Z = 0.84$, which corresponds to approximately 160 on QR. (b) $Z = -0.52$, which corresponds to approximately 147 on VR.

2.31 (a) $Z = 1.2 \rightarrow 0.1151$. (b) $Z = -1.28 \rightarrow 70.6^\circ$F or colder.

2.33 (a) $Z = 1.08 \rightarrow 0.1401$. (b) The answers are very close because only the units were changed. (The only reason why they are a little different is because 28°C is 82.4°F, not precisely 83°F.) (c) Since $IQR = Q3 - Q1$, we first need to find $Q3$ and $Q1$ and take the difference between the two. Remember that $Q3$ is the 75^{th} and $Q1$ is the 25^{th} percentile of a distribution. Q1 = 23.13, Q3 = 26.86, IQR = 26.86 - 23.13 = 3.73.

2.35 14/20 = 70% are within 1 SD of the mean. Within 2 SDs of the mean: 19/20 = 95%. Within 3 SDs of the mean: 20/20 = 100%. They follow this rule closely.

2.37 (a) We see the order of the categories and the relative frequencies in the bar plot. (b) There are no features that are apparent in the pie chart but not in the bar plot. (c) We usually prefer to use a bar plot as we can also see the relative frequencies of the categories in this graph.

2.39 The vertical locations at which the ideological groups break into the Yes, No, and Not Sure categories differ, which indicates that likelihood of supporting the DREAM act varies by political ideology. This suggests that the two variables may be dependent.

2.41 (a) (i) False. Instead of comparing counts, we should compare percentages of people in each group who suffered cardiovascular problems. (ii) True.

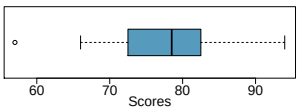
(iii) False. Association does not imply causation. We cannot infer a causal relationship based on an observational study. The difference from part (ii) is subtle. (iv) True.
(b) Proportion of all patients who had cardiovascular problems: $\frac{7,979}{227,571} \approx 0.035$
(c) The expected number of heart attacks in the rosiglitazone group, if having cardiovascular problems and treatment were independent, can be calculated as the number of patients in that group multiplied by the overall cardiovascular problem rate in the study: $67,593 * \frac{7,979}{227,571} \approx 2370$.
(d) (i) H_0: The treatment and cardiovascular problems are independent. They have no relationship, and the difference in incidence rates between the rosiglitazone and pioglitazone groups is due to chance. H_A: The treatment and cardiovascular problems are not independent. The difference in the incidence rates between the rosiglitazone and pioglitazone groups is not due to chance and rosiglitazone is associated with an increased risk of serious cardiovascular problems. (ii) A higher number of patients with cardiovascular problems than expected under the assumption of independence would provide support for the alternative hypothesis as this would suggest that rosiglitazone increases the risk of such problems. (iii) In the actual study, we observed 2,593 cardiovascular events in the rosiglitazone group. In the 1,000 simulations under the independence model, we observed somewhat less than 2,593 in every single simulation, which suggests that the actual results did not come from the independence model. That is, the variables do not appear to be independent, and we reject the independence model in favor of the alternative. The study's results provide convincing evidence that rosiglitazone is associated with an increased risk of cardiovascular problems.

2.43 (a) Decrease: the new score is smaller than the mean of the 24 previous scores. (b) Calculate a weighted mean. Use a weight of 24 for the old mean and 1 for the new mean: $(24 \times 74 + 1 \times 64)/(24 + 1) = 73.6$. (c) The new score is more than 1 standard deviation away from the previous mean, so increase.

2.45 No, we would expect this distribution to be right skewed. There are two reasons for this: (1) there is a natural boundary at 0 (it is not possible to watch less than 0 hours of TV), (2) the standard deviation of the distribution is very large compared to the mean.

2.47 The distribution of ages of best actress winners are right skewed with a median around 30 years. The distribution of ages of best actress winners is also right skewed, though less so, with a median around 40 years. The difference between the peaks of these distributions suggest that best actress winners are typically younger than best actor winners. The ages of best actress winners are more variable than the ages of best actor winners. There are potential outliers on the higher end of both of the distributions.

2.49



2.51 (a) $Z = \frac{5.5-7.44}{1.33} = -1.49; P(Z < -1.49) = 0.068$. Approximately 6.8% of the newborns were of low birth weight. (b) $Z = \frac{10-7.44}{1.33} = 1.925$. Using a lower bound of 2 and an upper bound of 5, we get $P(Z > 1.925) = 0.027$. Approximately 2.7% of the newborns weighed over 10 pounds. (c) Approximately 2.7% of the newborns weighed over 10 pounds. Because there were 23,419 of them, about $0.027 \times 23419 \approx 632$ weighed greater than 10 pounds. (d) Because we have the percentile, this is the inverse problem. To get the Z-score, use the inverse normal option with 0.90 to get $Z = 1.28$. Then solve for x in $1.28 = \frac{x-7.44}{1.33}$ to get $x = 9.15$. To be at the 90th percentile among this group, a newborn would have to weigh 9.15 pounds.

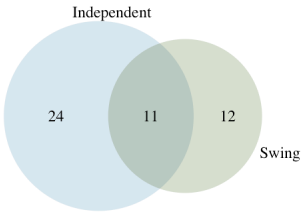
3 Probability

3.1 (a) False. These are independent trials. (b) False. There are red face cards. (c) True. A card cannot be both a face card and an ace.

3.3 (a) 10 tosses. Fewer tosses mean more variability in the sample fraction of heads, meaning there's a better chance of getting at least 60% heads. (b) 100 tosses. More flips means the observed proportion of heads would often be closer to the average, 0.50, and therefore also above 0.40. (c) 100 tosses. With more flips, the observed proportion of heads would often be closer to the average, 0.50. (d) 10 tosses. Fewer flips would increase variability in the fraction of tosses that are heads.

3.5 (a) $0.5^{10} = 0.00098$. (b) $0.5^{10} = 0.00098$. (c) $P(\text{at least one tails}) = 1 - P(\text{no tails}) = 1 - (0.5^{10}) \approx 1 - 0.001 = 0.999$.

3.7 (a) No, there are voters who are both independent and swing voters.
(b)



(c) Each Independent voter is either a swing voter or not. Since 35% of voters are Independents and 11% are both Independent and swing voters, the other 24% must not be swing voters. (d) 0.47. (e) 0.53. (f) P(Independent) × P(swing) $= 0.35 \times 0.23 = 0.08$, which does not equal P(Independent and swing) = 0.11, so the events are dependent.

3.9 (a) If the class is not graded on a curve, they are independent. If graded on a curve, then neither independent nor disjoint – unless the instructor will only give one A, which is a situation we will ignore in parts (b) and (c). (b) They are probably not independent: if you study together, your study habits would be related, which suggests your course performances are also related. (c) No. See the answer to part (a) when the course is not graded on a curve. More generally: if two things are unrelated (independent), then one occurring does not preclude the other from occurring.

3.11 (a) $0.16 + 0.09 = 0.25$. (b) $0.17 + 0.09 = 0.26$. (c) Assuming that the education level of the husband and wife are independent: $0.25 \times 0.26 = 0.065$. You might also notice we actually made a second assumption: that the decision to get married is unrelated to education level. (d) The husband/wife independence assumption is probably not reasonable, because people often marry another person with a comparable level of education. We will leave it to you to think about whether the second assumption noted in part (c) is reasonable.

3.13 (a) No, but we could if A and B are independent. (b-i) 0.21. (b-ii) 0.79. (b-iii) 0.3. (c) No, because $0.1 \neq 0.21$, where 0.21 was the value computed under independence from part (a). (d) 0.143.

3.15 (a) No, 0.18 of respondents fall into this combination. (b) $0.60 + 0.20 - 0.18 = 0.62$. (c) $0.18/0.20 = 0.9$. (d) $0.11/0.33 \approx 0.33$. (e) No, otherwise the answers to (c) and (d) would be the same. (f) $0.06/0.34 \approx 0.18$.

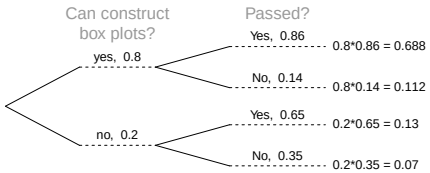
3.17 (a) No. There are 6 females who like Five Guys Burgers. (b) $162/248 = 0.65$. (c) $181/252 = 0.72$. (d) Under the assumption of a dating choices being independent of hamburger preference, which on the surface seems reasonable: $0.65 \times 0.72 = 0.468$. (e) $(252 + 6 - 1)/500 = 0.514$.

3.19 (a) 0.3. (b) 0.3. (c) 0.3. (d) $0.3 \times 0.3 = 0.09$. (e) Yes, the population that is being sampled from is identical in each draw.

3.21 (a) 2/9. (b) $3/9 = 1/3$. (c) $(3/10) \times (2/9) \approx 0.067$. (d) No. In this small population of marbles, removing one marble meaningfully changes the probability of what might be drawn next.

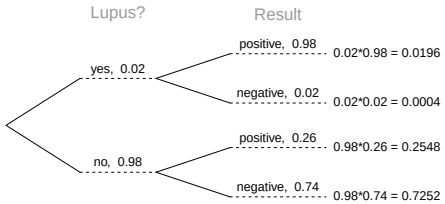
3.23 For 1 leggings (L) and 2 jeans (J), there are three possible orderings: LJJ, JLJ, and JJL. The probability for LJJ is $(5/24) \times (7/23) \times (6/22) = 0.0173$. The other two orderings have the same probability, and these three possible orderings are disjoint events. Final answer: 0.0519.

3.25 (a)



(b) 0.84

3.27 0.0714. Even when a patient tests positive for lupus, there is only a 7.14% chance that he actually has lupus. House may be right.



3.29 (a) P(pass) = 0.5, but it should be 0.16. (b) P(pass) = 0.2, instead of 0.16. (c) P(pass) = 0.17, instead of 0.16.

3.31 (a) Starting at row 3 of the random number table, we will read across the table two digits at a time. If the random number is between 00-15, the car will fail the pollution test. If the number is between 16-99, the car will pass the test. (Answers may vary.) (b) Fleet 1: 18-52-97-32-85-95-29 → P-P-P-P-P-P-P → fleet passes
Fleet 2: 14-96-06-67-17-49-59 → F-P-F-P-P-P-P → fleet fails
Fleet 3: 05-33-67-97-58-11-81 → F-P-P-P-P-F-P → fleet fails
Fleet 4: 23-81-83-21-71-08-50 → P-P-P-P-P-F-P → fleet fails
Fleet 5: 82-84-39-31-83-14-34 → P-P-P-P-P-F-P → fleet fails (c) 4 / 5 = 0.80

3.33 (a) 13. (b) No, these 27 students are not a random sample from the university's student population. For example, it might be argued that the proportion of smokers among students who go to the gym at 9 am on a Saturday morning would be lower than the proportion of smokers in the university as a whole.

3.35 (a) E(X) = 3.59. SD(X) = 9.64. (b) E(X) = -1.41. SD(X) = 9.64. (c) No, the expected net profit is negative, so on average you expect to lose money.

3.37 5% increase in value.

3.39 E = -0.0526. SD = 0.9986.

3.41 (a) Let X represent the amount of lemonade in the pitcher, Y represent the amount of lemonade in a glass, and W represent the amount left over after. Then, $\mu_W = E(X - Y) = 64 - 12 = 52$ (b) $\sigma_W = \sqrt{SD(X)^2 + SD(Y)^2} = \sqrt{1.732^2 + 1^2} \approx \sqrt{4} = 2$ (c) $P(W > 50) = P\left(Z > \frac{50-52}{2}\right) = P(Z > -1) = 1 - 0.1587 = 0.8413$

3.43 (a) The combined scores follow a normal distribution with $\mu_{combined} = 304$ and $\sigma_{combined} = 10.38$. Then, P(combined score > 320) is approximately 0.06. (b) Z=1.28 (using calculator or table). Then we set $1.28 = \frac{x-304}{10.38}$ and find $x \approx 317$.

3.45 (a) No. The cards are not independent. For example, if the first card is an ace of clubs, that implies the second card cannot be an ace of clubs. Additionally, there are many possible categories, which would need to be simplified. (b) No. There are six events under consideration. The Bernoulli distribution allows for only two events or categories. Note that rolling a die could be a Bernoulli trial if we simply to two events, e.g. rolling a 6 and not rolling a 6, though specifying such details would be necessary.

3.47 (a) $0.875^2 \times 0.125 = 0.096$. (b) $\mu = 8$, $\sigma = 7.48$.

3.49 If p is the probability of a success, then the mean of a Bernoulli random variable X is given by
$\mu = E[X] = P(X = 0) \times 0 + P(X = 1) \times 1$
$= (1 - p) \times 0 + p \times 1 = 0 + p = p$

3.51 (a) $\binom{5}{1} = 5$. (b) $\binom{5}{4} = 5$. (c) $\binom{5}{3} = 10$. (d) $\binom{5}{3} + \binom{5}{4} + \binom{5}{5} = 10 + 5 + 1 = 16$.

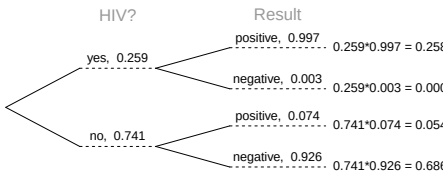
3.53 (a) Yes. The conditions are satisfied: independence, fixed number of trials, either success or failure for each trial, and probability of success being constant across trials. (b) 0.200. (c) 0.200. (d) $0.0024 + 0.0284 + 0.1323 = 0.1631$. (e) $1 - 0.0024 = 0.9976$.

3.55 (a) $1 - 0.75^3 = 0.5781$. (b) 0.1406. (c) 0.4219. (d) $1 - 0.25^3 = 0.9844$.

3.57 (a) $\mu = 35$, $\sigma = 3.24$. (b) Yes. $Z = 3.09$. Since 45 is more than 2 standard deviations from the mean, it would be considered unusual. Note that the normal model is not required to apply this rule of thumb. (c) Using a normal model: 0.0010. This does indeed appear to be an unusual observation. If using a normal model with a 0.5 correction, the probability would be calculated as 0.0017.

3.59 (a) Invalid. Sum is greater than 1. (b) Valid. Probabilities are between 0 and 1, and they sum to 1. In this class, every student gets a C. (c) Invalid. Sum is less than 1. (d) Invalid. There is a negative probability. (e) Valid. Probabilities are between 0 and 1, and they sum to 1. (f) Invalid. There is a negative probability.

3.61 0.8247.



3.63 (a) E = \$3.90. SD = \$0.34.
(b) E = \$27.30. SD = \$0.89.

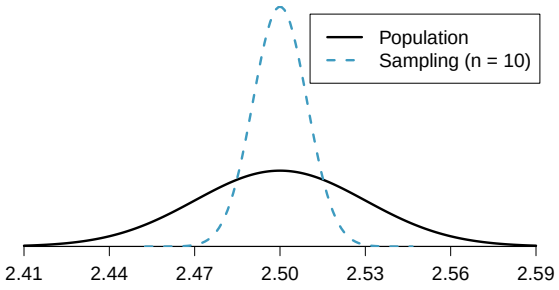
3.65 Want to find the probability that there will be 1,786 or more enrollees. Using the normal approximation, with $\mu = np = 2,500 \times 0.7 = 1750$ and $\sigma = \sqrt{np(1-p)} = \sqrt{2,500 \times 0.7 \times 0.3} \approx 23$, $Z = 1.61$, and $P(Z > 1.61) = 0.0537$. With a 0.5 correction: 0.0559.

?? (a) The distribution is unimodal and strongly right skewed with a median between 5 and 10 years old. Ages range from 0 to slightly over 50 years old, and the middle 50% of the distribution is roughly between 5 and 15 years old. There are potential outliers on the higher end. (b) When the sample size is small, the sampling distribution is right skewed, just like the population distribution. As the sample size increases, the sampling distribution gets more unimodal, symmetric, and approaches normality. The variability also decreases. This is consistent with the Central Limit Theorem.

?? (a) Right skewed. There is a long tail on the higher end of the distribution but a much shorter tail on the lower end. (b) Less than, as the median would be less than the mean in a right skewed distribution. (c) We should not. (d) Even though the population distribution is not normal, the conditions for inference are reasonably satisfied, with the possible exception of skew. If the skew isn't very strong (we should ask to see the data), then we can use the Central Limit Theorem to estimate this probability. For now, we'll assume the skew isn't very strong, though the description suggests it is at least moderate to strong. Use $N(1.3, SD_{\bar{x}} = 0.3/\sqrt{60})$: $Z = 2.58 \rightarrow 0.0049$. (e) It would decrease it by a factor of $\sqrt{2}$.

?? The centers are the same in each plot, and each data set is from a nearly normal distribution (see Section **??**), though the histograms may not look very normal since each represents only 100 data points. The only way to tell which plot corresponds to which scenario is to examine the variability of each distribution. Plot B is the most variable, followed by Plot A, then Plot C. This means Plot B will correspond to the original data, Plot A to the sample means with size 5, and Plot C to the sample means with size 25.

?? (a) $Z = -3.33 \rightarrow 0.0004$. (b) The population SD is known and the data are nearly normal, so the sample mean will be nearly normal with distribution $N(\mu, \sigma/\sqrt{n})$, i.e. $N(2.5, 0.0095)$. (c) $Z = -10.54 \rightarrow \approx 0$. (d) See below:



(e) We could not estimate (a) without a nearly normal population distribution. We also could not estimate (c) since the sample size is not sufficient to yield a nearly normal sampling distribution if the population distribution is not nearly normal.

?? (a) We cannot use the normal model for this calculation, but we can use the histogram. About 500 songs are shown to be longer than 5 minutes, so the probability is about $500/3000 = 0.167$. (b) Two different answers are reasonable. *Option 1* Since the population distribution is only slightly skewed to the right, even a small sample size will yield a nearly normal sampling distribution. We also know that the songs are sampled randomly and the sample size is less than 10% of the population, so the length of one song in the sample is independent of another. We are looking for the probability that the total length of 15 songs is more than 60 minutes, which means that the average song should last at least $60/15 = 4$ minutes. Using $SD_{\bar{x}} = 1.63/\sqrt{15}$, $Z = 1.31 \rightarrow 0.0951$. *Option 2* Since the population distribution is not normal, a small sample size may not be sufficient to yield a nearly normal sampling distribution. Therefore, we cannot estimate the probability using the tools we have learned so far. (c) We can now be confident that the conditions are satisfied. $Z = 0.92 \rightarrow 0.1788$.

?? (a) $SD_{\bar{x}} = \frac{25}{\sqrt{75}} = 2.89$. (b) $Z = 1.73$, which indicates that the two values are not unusually distant from each other when accounting for the uncertainty in John's point estimate.

?? (a) Each observation in each of the distributions represents the sample proportion ($\hat{p}$) from samples of size $n = 20$, $n = 100$, and $n = 500$, respectively. (b) The centers for all three distributions are at 0.95, the true population parameter. When n is small, the distribution is skewed to the left and not smooth. As n increases, the variability of the distribution (standard deviation) decreases, and the shape of the distribution becomes more unimodal and symmetric.

?? (a) $SD_{\hat{p}} = \sqrt{p(1-p)/n} = 0.0707$. This describes the typical distance that the sample proportion will deviate from the true proportion, $p = 0.5$. (b) $\hat{p}$ approximately follows $N(0.5, 0.0707)$. $Z = (0.55 - 0.50)/0.0707 \approx 0.71$. This corresponds to an upper tail of about 0.2389. That is, $P(\hat{p} > 0.55) \approx 0.24$.

?? (a) First we need to check that the necessary conditions are met. There are $200 \times 0.08 = 16$ expected successes and $200 \times (1-0.08) = 184$ expected failures, therefore the success-failure condition is met. Then the binomial distribution can be approximated by $N(\mu = 16, \sigma = 3.84)$. $P(X < 12) = P(Z < -1.04) = 0.1492$. (b) Since the success-failure condition is met the sampling distribution of $\hat{p} \sim N(\mu = 0.08, \sigma = 0.0192)$. $P(\hat{p} < 0.06) = P(Z < -1.04) = 0.1492$. (c) As expected, the two answers are the same.

?? 0 wins (-\$3): 0.1458. 1 win (-\$1): 0.3936. 2 wins (+\$1): 0.3543. 3 wins (+\$3): 0.1063.

?? (a) $Z = 0.67$. (b) $\mu = \$1650$, $x = \$1800$. (c) $0.67 = \frac{1800-1650}{\sigma} \rightarrow \sigma = \$223.88$.

?? (a) $(1 - 0.471)^2 \times 0.471 = 0.1318$. (b) $0.471^3 = 0.1045$. (c) $\mu = 1/0.471 = 2.12$, $\sigma = \sqrt{2.38} = 1.54$. (d) $\mu = 1/0.30 = 3.33$, $\sigma = 2.79$. (e) When p is smaller, the event is rarer, meaning the expected number of trials before a success and the standard deviation of the waiting time are higher.

?? $Z = 1.56$, $P(Z > 1.56) = 0.0594$, i.e. 6%.

?? (a) $Z = 0.73$, $P(Z > 0.73) = 0.2327$. (b) If you are bidding on only one auction and set a low maximum bid price, someone will probably outbid you. If you set a high maximum bid price, you may win the auction but pay more than is necessary. If bidding on more than one auction, and you set your maximum bid price very low, you probably won't win any of the auctions. However, if the maximum bid price is even modestly high, you are likely to win multiple auctions. (c) An answer roughly equal to the 10th percentile would be reasonable. Regrettably, no percentile cutoff point guarantees beyond any possible event that you win at least one auction. However, you may pick a higher percentile if you want to be more sure of winning an auction. (d) Answers will vary a little but should correspond to the answer in part (c). We use the 10^{th} percentile: $Z = -1.28 \rightarrow \$69.80$.

?? (a) $Z = 3.5$, upper tail is 0.0002. (More precise value: 0.000233, but we'll use 0.0002 for the calculations here.)
(b) $0.0002 \times 2000 = 0.4$. We would expect about 0.4 10 year olds who are 76 inches or taller to show up.
(c) $\binom{2000}{0}(0.0002)^0(1 - 0.0002)^{2000} = 0.67029$.
(d) $\frac{0.4^0 \times e^{-0.4}}{0!} = \frac{1 \times e^{-0.4}}{1} = 0.67032$.

?? This is the same as checking that the average bag weight of the 10 bags is greater than 46 lbs. $SD_{\bar{x}} = \frac{3.2}{\sqrt{10}} = 1.012$; $z = \frac{46-45}{1.012} = 0.988$; $P(z > 0.988) = 0.162 = 16.2\%$.

?? First we need to check that the necessary conditions are met. There are $100 \times 0.389 = 38.9$ expected successes and $100 \times (1 - 0.389) = 61.1$ expected failures, therefore the success-failure condition is met. Calculate using either (1) the normal approximation to the binomial distribution or (2) the sampling distribution of $\hat{p}$. (1) The binomial distribution can be approximated by $N(\mu = 0.389, \sigma = 4.88)$. $P(X \geq 35) = P(Z > -0.80) = 1 - 0.2119 = 0.7881$. (2) The sampling distribution of $\hat{p} \sim N(\mu = 0.389, \sigma = 0.0488)$. $P(\hat{p} > 0.35) = P(Z > -0.8) = 0.7881$.

5 Foundations for inference

?? (a) Mean. Each student reports a numerical value: a number of hours. (b) Mean. Each student reports a number, which is a percentage, and we can average over these percentages. (c) Proportion. Each student reports Yes or No, so this is a categorical variable and we use a proportion. (d) Mean. Each student reports a number, which is a percentage like in part (b). (e) Proportion. Each student reports whether or not s/he expects to get a job, so this is a categorical variable and we use a proportion.

?? (a) The sample is from all computer chips manufactured at the factory during the week of production. We might be tempted to generalize the population to represent all weeks, but we should exercise caution here since the rate of defects may change over time. (b) The fraction of computer chips manufactured at the factory during the week of production that had defects. (c) Estimate the parameter using the data: $\hat{p} = \frac{27}{212} = 0.127$. (d) *Standard error* (or SE). (e) Compute the SE using $\hat{p} = 0.127$ in place of p: $SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.127(1-0.127)}{212}} = 0.023$. (f) The standard error is the standard deviation of $\hat{p}$. A value of 0.10 would be about one standard error away from the observed value, which would not represent a very uncommon deviation. (Usually beyond about 2 standard errors is a good rule of thumb.) The engineer should not be surprised. (g) Recomputed standard error using $p = 0.1$: $SE = \sqrt{\frac{0.1(1-0.1)}{212}} = 0.021$. This value isn't very different, which is typical when the standard error is computed using relatively similar proportions (and even sometimes when those proportions are quite different!).

?? (a) Sampling distribution. (b) If the population proportion is in the 5-30% range, the success-failure condition would be satisfied and the sampling distribution would be symmetric. (c) We use the formula for the standard error: $SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08(1-0.08)}{800}} = 0.0096$. (d) Standard error. (e) The distribution will tend to be more variable when we have fewer observations per sample.

?? Recall that the general formula is *point estimate* $\pm\ z^\star \times SE$. First, identify the three different values. The point estimate is 45%, $z^\star = 1.96$ for a 95% confidence level, and $SE = 1.2\%$. Then, plug the values into the formula: $45\% \pm 1.96 \times 1.2\% \rightarrow (42.6\%, 47.4\%)$ We are 95% confident that the proportion of US adults who live with one or more chronic conditions is between 42.6% and 47.4%.

?? (a) False. Confidence intervals provide a range of plausible values, and sometimes the truth is missed. A 95% confidence interval "misses" about 5% of the time. (b) True. Notice that the description focuses on the true population value. (c) True. If we examine the 95% confidence interval computed in Exercise **??**, we can see that 50% is not included in this interval. This means that in a hypothesis test, we would reject the null hypothesis that the proportion is 0.5. (d) False. The standard error describes the uncertainty in the overall estimate from natural fluctuations due to randomness, not the uncertainty corresponding to individuals' responses.

?? (a) False. Inference is made on the population parameter, not the point estimate. The point estimate is always in the confidence interval. (b) True. (c) False. The confidence interval is not about a sample mean. (d) False. To be more confident that we capture the parameter, we need a wider interval. Think about needing a bigger net to be more sure of catching a fish in a murky lake. (e) True. Optional explanation: This is true since the normal model was used to model the sample mean. The margin of error is half the width of the interval, and the sample mean is the midpoint of the interval. (f) False. In the calculation of the standard error, we divide the standard deviation by the square root of the sample size. To cut the SE (or margin of error) in half, we would need to sample $2^2 = 4$ times the number of people in the initial sample.

?? (a) $H_0 : p = 0.5$ (Neither a majority nor minority of students' grades improved) $H_A : p \neq 0.5$ (Either a majority or a minority of students' grades improved) (b) $H_0 : \mu = 15$ (The average amount of company time each employee spends not working is 15 minutes for March Madness.) $H_A : \mu \neq 15$ (The average amount of company time each employee spends not working is different than 15 minutes for March Madness.)

?? (1) The hypotheses should be about the population proportion (p), not the sample proportion. (2) The null hypothesis should have an equal sign. (3) The alternative hypothesis should have a not-equals sign, and (4) it should reference the null value, $p_0 = 0.6$, not the observed sample proportion. The correct way to set up these hypotheses is: $H_0 : p = 0.6$ and $H_A : p \neq 0.6$.

?? (a) This claim is reasonable, since the entire interval lies above 50%. (b) The value of 70% lies outside of the interval, so we have convincing evidence that the researcher's conjecture is wrong. (c) A 90% confidence interval will be narrower than a 95% confidence interval. Even without calculating the interval, we can tell that 70% would not fall in the interval, and we would reject the researcher's conjecture based on a 90% confidence level as well.

?? (i) Set up hypotheses. H_0: $p = 0.5$, H_A: $p \neq 0.5$. We will use a significance level of $\alpha = 0.05$. (ii) Check conditions: simple random sample gets us independence, and the success-failure conditions is satisfied since $0.42 \times 1000 = 420$ and $(1 - 0.42) \times 1000 = 580$ are both at least 10. (iii) Next, we calculate: $SE = \sqrt{0.5(1-0.5)/1000} = 0.016$. $Z = \frac{0.42-0.5}{0.016} = -5$, which has a one-tail area of about 0.0000003, so the p-value is twice this one-tail area at 0.0000006. (iv) Make a conclusion: Because the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that the fraction of US adults who believe raising the minimum wage will help the economy is not 50%. Because the observed value is less than 50% and we have rejected the null hypothesis, we can conclude that this belief is held by fewer than 50% of US adults. (For reference, the survey also explores support for changing the minimum wage, which is a different question than if it will help the economy.)

?? If the p-value is 0.05, this means the test statistic would be either $Z = -1.96$ or $Z = 1.96$. We'll show the calculations for $Z = 1.96$. Standard error: $SE = \sqrt{0.3(1-0.3)/90} = 0.048$. Finally, set up the test statistic formula and solve for $\hat{p}$: $1.96 = \frac{\hat{p}-0.3}{0.048} \rightarrow \hat{p} = 0.394$ Alternatively, if $Z = -1.96$ was used: $\hat{p} = 0.206$.

?? (a) H_0: Anti-depressants do not affect the symptoms of Fibromyalgia. H_A: Anti-depressants do affect the symptoms of Fibromyalgia (either helping or harming). (b) Concluding that anti-depressants either help or worsen Fibromyalgia symptoms when they actually do neither. (c) Concluding that anti-depressants do not affect Fibromyalgia symptoms when they actually do.

?? (a) We are 95% confident that Americans spend an average of 1.38 to 1.92 hours per day relaxing or pursuing activities they enjoy. (b) Their confidence level must be higher as the width of the confidence interval increases as the confidence level increases. (c) The new margin of error will be smaller, since as the sample size increases, the standard error decreases, which will decrease the margin of error.

?? (a) H_0: The restaurant meets food safety and sanitation regulations. H_A: The restaurant does not meet food safety and sanitation regulations. (b) The food safety inspector concludes that the restaurant does not meet food safety and sanitation regulations and shuts down the restaurant when the restaurant is actually safe. (c) The food safety inspector concludes that the restaurant meets food safety and sanitation regulations and the restaurant stays open when the restaurant is actually not safe. (d) A Type 1 Error may be more problematic for the restaurant owner since his restaurant gets shut down even though it meets the food safety and sanitation regulations. (e) A Type 2 Error may be more problematic for diners since the restaurant deemed safe by the inspector is actually not. (f) Strong evidence. Diners would rather a restaurant that meet the regulations get shut down than a restaurant that doesn't meet the regulations not get shut down.

?? (a) $H_0 : p_{unemp} = p_{underemp}$: The proportions of unemployed and underemployed people who are having relationship problems are equal. $H_A : p_{unemp} \neq punderemp$: The proportions of unemployed and underemployed people who are having relationship problems are different. (b) If in fact the two population proportions are equal, the probability of observing at least a 2% difference between the sample proportions is approximately 0.35. Since this is a high probability we fail to reject the null hypothesis. The data do not provide convincing evidence that the proportion of of unemployed and underemployed people who are having relationship problems are different.

?? Because 130 is inside the confidence interval, we do not have convincing evidence that the true average is any different than what the nutrition label suggests.

?? True. If the sample size gets ever larger, then the standard error will become ever smaller. Eventually, when the sample size is large enough and the standard error is tiny, we can find statistically significant yet very small differences between the null value and point estimate (assuming they are not exactly equal).

?? (a) In effect, we're checking whether men are paid more than women (or vice-versa), and we'd expect these outcomes with either chance under the null hypothesis:

$$H_0 : p = 0.5 \qquad H_A : p \neq 0.5$$

We'll use p to represent the fraction of cases where men are paid more than women.
(b) Below is the completion of the hypothesis test.

- There isn't a good way to check independence here since the jobs are not a simple random sample. However, independence doesn't seem unreasonable, since the individuals in each job are different from each other. The success-failure condition is met since we check it using the null proportion: $p_0 n = (1 - p_0)n = 10.5$ is greater than 10.
- We can compute the sample proportion, SE, and test statistic:

$$\hat{p} = 19/21 = 0.905$$
$$SE = \sqrt{\frac{0.5 \times (1 - 0.5)}{21}} = 0.109$$
$$Z = \frac{0.905 - 0.5}{0.109} = 3.72$$

  The test statistic Z corresponds to an upper tail area of about 0.0001, so the p-value is 2 times this value: 0.0002.
- Because the p-value is smaller than 0.05, we reject the notion that all these gender pay disparities are due to chance. Because we observe that men are paid more in a higher proportion of cases and we have rejected H_0, we can conclude that men are being paid higher amounts in ways not explainable by chance alone.

If you're curious for more info around this topic, including a discussion about adjusting for additional factors that affect pay, please see the following video by Healthcare Triage: youtu.be/aVhgKSULNQA.

6 Inference for categorical data

?? (a) False. Doesn't satisfy success-failure condition. (b) True. The success-failure condition is not satisfied. In most samples we would expect $\hat{p}$ to be close to 0.08, the true population proportion. While $\hat{p}$ can be much above 0.08, it is bound below by 0, suggesting it would take on a right skewed shape. Plotting the sampling distribution would confirm this suspicion. (c) False. $SE_{\hat{p}} = 0.0243$, and $\hat{p} = 0.12$ is only $\frac{0.12-0.08}{0.0243} = 1.65$ SEs away from the mean, which would not be considered unusual. (d) True. $\hat{p} = 0.12$ is 2.32 standard errors away from the mean, which is often considered unusual. (e) False. Decreases the SE by a factor of $1/\sqrt{2}$.

?? (a) True. See the reasoning of 6.1(b). (b) True. We take the square root of the sample size in the SE formula. (c) True. The independence and success-failure conditions are satisfied. (d) True. The independence and success-failure conditions are satisfied.

?? (a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion. (b) True. 95% CI: 82% ± 2%. (c) True. By the definition of the confidence level. (d) True. Quadrupling the sample size decreases the SE and ME by a factor of $1/\sqrt{4}$. (e) True. The 95% CI is entirely above 50%.

?? With a random sample, independence is satisfied. The success-failure condition is also satisfied. $ME = z^\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56\times0.44}{600}} = 0.0397 \approx 4\%$

?? (a) No. The sample only represents students who took the SAT, and this was also an online survey. (b) (0.5289, 0.5711). We are 90% confident that 53% to 57% of high school seniors who took the SAT are fairly certain that they will participate in a study abroad program in college. (c) 90% of such random samples would produce a 90% confidence interval that includes the true proportion. (d) Yes. The interval lies entirely above 50%.

?? (a) We want to check for a majority (or minority), so we use the following hypotheses:

$$H_0 : p = 0.5 \qquad H_A : p \neq 0.5$$

We have a sample proportion of $\hat{p} = 0.55$ and a sample size of $n = 617$ independents.
Since this is a random sample, independence is satisfied. The success-failure condition is also satisfied: 617×0.5 and $617 \times (1 - 0.5)$ are both at least 10 (we use the null proportion $p_0 = 0.5$ for this check in a one-proportion hypothesis test).
Therefore, we can model $\hat{p}$ using a normal distribution with a standard error of

$$SE = \sqrt{\frac{p(1-p)}{n}} = 0.02$$

(We use the null proportion $p_0 = 0.5$ to compute the standard error for a one-proportion hypothesis test.)
Next, we compute the test statistic:

$$Z = \frac{0.55 - 0.5}{0.02} = 2.5$$

This yields a one-tail area of 0.0062, and a p-value of $2 \times 0.0062 = 0.0124$.
Because the p-value is smaller than 0.05, we reject the null hypothesis. We have strong evidence that the support is different from 0.5, and since the data provide a point estimate above 0.5, we have strong evidence to support this claim by the TV pundit.
(b) No. Generally we expect a hypothesis test and a confidence interval to align, so we would expect the confidence interval to show a range of plausible values entirely above 0.5. However, if the confidence level is misaligned (e.g. a 99% confidence level and a $\alpha = 0.05$ significance level), then this is no longer generally true.

?? (a) $H_0 : p = 0.5$. $H_A : p \neq 0.5$. Independence (random sample) is satisfied, as is the success-failure conditions (using $p_0 = 0.5$, we expect 40 successes and 40 failures). $Z = 2.91 \rightarrow$ the one tail area is 0.0018, so the p-value is 0.0036. Since the p-value < 0.05, we reject the null hypothesis. Since we rejected H_0 and the point estimate suggests people are better than random guessing, we can conclude the rate of correctly identifying a soda for these people is significantly better than just by random guessing.
(b) If in fact people cannot tell the difference between diet and regular soda and they were randomly guessing, the probability of getting a random sample of 80 people where 53 or more identify a soda correctly (or 53 or more identify a soda incorrectly) would be 0.0036.

?? Since a sample proportion ($\hat{p} = 0.55$) is available, we use this for the sample size calculations. The margin of error for a 90% confidence interval is $1.65 \times SE = 1.65 \times \sqrt{\frac{p(1-p)}{n}}$. We want this to be less than 0.01, where we use $\hat{p}$ in place of p:

$$1.65 \times \sqrt{\frac{0.55(1-0.55)}{n}} \leq 0.01$$
$$1.65^2 \frac{0.55(1-0.55)}{0.01^2} \leq n$$

From this, we get that n must be at least 6739.

?? This is not a randomized experiment, and it is unclear whether people would be affected by the behavior of their peers. That is, independence may not hold. Additionally, there are only 5 interventions under the provocative scenario, so the success-failure condition does not hold. Even if we consider a hypothesis test where we pool the proportions, the success-failure condition will not be satisfied. Since one condition is questionable and the other is not satisfied, the difference in sample proportions will not follow a nearly normal distribution.

?? (a) False. The entire confidence interval is above 0. (b) True. (c) True. (d) True. (e) False. It is simply the negated and reordered values: (-0.06,-0.02).

?? (a) Standard error:

$$SE = \sqrt{\frac{0.79(1-0.79)}{347} + \frac{0.55(1-0.55)}{617}} = 0.03$$

Using $z^\star = 1.96$, we get:

$$0.79 - 0.55 \pm 1.96 \times 0.03 \rightarrow (0.181, 0.299)$$

We are 95% confident that the proportion of Democrats who support the plan is 18.1% to 29.9% higher than the proportion of Independents who support the plan. (b) True.

?? (a) College grads: 23.7%. Non-college grads: 33.7%. (b) Let p_{CG} and p_{NCG} represent the proportion of college graduates and non-college graduates who responded "do not know". $H_0 : p_{CG} = p_{NCG}$. $H_A : p_{CG} \neq p_{NCG}$. Independence is satisfied (random sample), and the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 235/827 = 0.284$), is also satisfied. $Z = -3.18 \rightarrow$ p-value $= 0.0014$. Since the p-value is very small, we reject H_0. The data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates. The data also indicate that fewer college grads say they "do not know" than non-college grads (i.e. the data indicate the direction after we reject H_0).

?? (a) College grads: 35.2%. Non-college grads: 33.9%. (b) Let p_{CG} and p_{NCG} represent the proportion of college graduates and non-college grads who support offshore drilling. $H_0 : p_{CG} = p_{NCG}$. $H_A : p_{CG} \neq p_{NCG}$. Independence is satisfied (random sample), and the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 286/827 = 0.346$), is also satisfied. $Z = 0.39 \rightarrow$ p-value $= 0.6966$. Since the p-value $> \alpha$ (0.05), we fail to reject H_0. The data do not provide strong evidence of a difference between the proportions of college graduates and non-college graduates who support off-shore drilling in California.

?? Subscript $_C$ means control group. Subscript $_T$ means truck drivers. $H_0 : p_C = p_T$. $H_A : p_C \neq p_T$. Independence is satisfied (random samples), as is the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 70/495 = 0.141$). $Z = -1.65 \rightarrow$ p-value $= 0.0989$. Since the p-value is high (default to alpha $= 0.05$), we fail to reject H_0. The data do not provide strong evidence that the rates of sleep deprivation are different for non-transportation workers and truck drivers.

?? (a) Summary of the study:

		Virol. failure		
		Yes	No	Total
Treatment	Nevaripine	26	94	120
	Lopinavir	10	110	120
	Total	36	204	240

(b) $H_0 : p_N = p_L$. There is no difference in virologic failure rates between the Nevaripine and Lopinavir groups. $H_A : p_N \neq p_L$. There is some difference in virologic failure rates between the Nevaripine and Lopinavir groups. (c) Random assignment was used, so the observations in each group are independent. If the patients in the study are representative of those in the general population (something impossible to check with the given information), then we can also confidently generalize the findings to the population. The success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 36/240 = 0.15$), is satisfied. $Z = 2.89 \rightarrow$ p-value $= 0.0039$. Since the p-value is low, we reject H_0. There is strong evidence of a difference in virologic failure rates between the Nevaripine and Lopinavir groups. Treatment and virologic failure do not appear to be independent.

?? (a) False. The chi-square distribution has one parameter called degrees of freedom. (b) True. (c) True. (d) False. As the degrees of freedom increases, the shape of the chi-square distribution becomes more symmetric.

?? (a) H_0: The distribution of the format of the book used by the students follows the professor's predictions. H_A: The distribution of the format of the book used by the students does not follow the professor's predictions. (b) $E_{hard\ copy} = 126 \times 0.60 = 75.6$. $E_{print} = 126 \times 0.25 = 31.5$. $E_{online} = 126 \times 0.15 = 18.9$. (c) Independence: The sample is not random. However, if the professor has reason to believe that the proportions are stable from one term to the next and students are not affecting each other's study habits, independence is probably reasonable. Sample size: All expected counts are at least 5. (d) $\chi^2 = 2.32$, $df = 2$, p-value = 0.313. (e) Since the p-value is large, we fail to reject H_0. The data do not provide strong evidence indicating the professor's predictions were statistically inaccurate.

?? (a) Two-way table:

	Quit		
Treatment	Yes	No	Total
Patch + support group	40	110	150
Only patch	30	120	150
Total	70	230	300

(b-i) $E_{row_1,col_1} = \frac{(row\ 1\ total) \times (col\ 1\ total)}{table\ total} = 35$. This is lower than the observed value.
(b-ii) $E_{row_2,col_2} = \frac{(row\ 2\ total) \times (col\ 2\ total)}{table\ total} = 115$. This is lower than the observed value.

?? H_0: The opinion of college grads and non-grads is not different on the topic of drilling for oil and natural gas off the coast of California. H_A: Opinions regarding the drilling for oil and natural gas off the coast of California has an association with earning a college degree.

$$E_{row\ 1,col\ 1} = 151.5 \qquad E_{row\ 1,col\ 2} = 134.5$$
$$E_{row\ 2,col\ 1} = 162.1 \qquad E_{row\ 2,col\ 2} = 143.9$$
$$E_{row\ 3,col\ 1} = 124.5 \qquad E_{row\ 3,col\ 2} = 110.5$$

Independence: The samples are both random, unrelated, and from less than 10% of the population, so independence between observations is reasonable. Sample size: All expected counts are at least 5. $\chi^2 = 11.47$, $df = 2 \rightarrow$ p-value = 0.003. Since the p-value $< \alpha$, we reject H_0. There is strong evidence that there is an association between support for offshore drilling and having a college degree.

?? No. The samples at the beginning and at the end of the semester are not independent since the survey is conducted on the same students.

?? (a) H_0: The age of Los Angeles residents is independent of shipping carrier preference variable. H_A: The age of Los Angeles residents is associated with the shipping carrier preference variable. (b) The conditions are not satisfied since some expected counts are below 5.

?? (a) Independence is satisfied (random sample), as is the success-failure condition (40 smokers, 160 non-smokers). The 95% CI: (0.145, 0.255). We are 95% confident that 14.5% to 25.5% of all students at this university smoke. (b) We want $z^\star SE$ to be no larger than 0.02 for a 95% confidence level. We use $z^\star = 1.96$ and plug in the point estimate $\hat{p} = 0.2$ within the SE formula: $1.96\sqrt{0.2(1-0.2)/n} \leq 0.02$. The sample size n should be at least 1,537.

?? (a) Proportion of graduates from this university who found a job within one year of graduating. $\hat{p} = 348/400 = 0.87$. (b) This is a random sample, so the observations are independent. Success-failure condition is satisfied: 348 successes, 52 failures, both well above 10. (c) (0.8371, 0.9029). We are 95% confident that approximately 84% to 90% of graduates from this university found a job within one year of completing their undergraduate degree. (d) 95% of such random samples would produce a 95% confidence interval that includes the true proportion of students at this university who found a job within one year of graduating from college. (e) (0.8267, 0.9133). Similar interpretation as before. (f) 99% CI is wider, as we are more confident that the true proportion is within the interval and so need to cover a wider range.

?? Use a chi-squared goodness of fit test. H_0: Each option is equally likely. H_A: Some options are preferred over others. Total sample size: 99. Expected counts: (1/3) * 99 = 33 for each option. These are all above 5, so conditions are satisfied. $df = 3 - 1 = 2$ and $\chi^2 = \frac{(43-33)^2}{33} + \frac{(21-33)^2}{33} + \frac{(35-33)^2}{33} = 7.52 \rightarrow$ p-value = 0.023. Since the p-value is less than 5%, we reject H_0. The data provide convincing evidence that some options are preferred over others.

?? (a) $H_0 : p = 0.38$. $H_A : p \neq 0.38$. Independence (random sample) and the success-failure condition are satisfied. $Z = -20.5 \rightarrow$ p-value ≈ 0. Since the p-value is very small, we reject H_0. The data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%, and the data indicate that the proportion is lower in the US. (b) If in fact 38% of Americans used their cell phones as a primary access point to the internet, the probability of obtaining a random sample of 2,254 Americans where 17% or less or 59% or more use their only their cell phones to access the internet would be approximately 0. (c) (0.1545, 0.1855). We are 95% confident that approximately 15.5% to 18.6% of all Americans primarily use their cell phones to browse the internet.

7 Inference for numerical data

?? (a) $df = 6 - 1 = 5$, $t^\star_5 = 2.02$ (column with two tails of 0.10, row with $df = 5$). (b) $df = 21 - 1 = 20$, $t^\star_{20} = 2.53$ (column with two tails of 0.02, row with $df = 20$). (c) $df = 28$, $t^\star_{28} = 2.05$. (d) $df = 11$, $t^\star_{11} = 3.11$.

?? (a) 0.085, do not reject H_0. (b) 0.003, reject H_0. (c) 0.438, do not reject H_0. (d) 0.042, reject H_0.

?? The mean is the midpoint: $\bar{x} = 20$. Identify the margin of error: $ME = 1.015$, then use $t^\star_{35} = 2.03$ and $SE = s/\sqrt{n}$ in the formula for margin of error to identify $s = 3$.

?? (a) H_0: $\mu = 8$ (New Yorkers sleep 8 hrs per night on average.) H_A: $\mu \neq 8$ (New Yorkers sleep less or more than 8 hrs per night on average.) (b) Independence: The sample is random. The min/max suggest there are no concerning outliers. $T = -1.75$. $df = 25 - 1 = 24$. (c) p-value = 0.093. If in fact the true population mean of the amount New Yorkers sleep per night was 8 hours, the probability of getting a random sample of 25 New Yorkers where the average amount of sleep is 7.73 hours per night or less (or 8.27 hours or more) is 0.093. (d) Since p-value > 0.05, do not reject H_0. The data do not provide strong evidence that New Yorkers sleep more or less than 8 hours per night on average. (e) Yes, since we did not rejected H_0.

?? T is either -2.09 or 2.09. Then $\bar{x}$ is one of the following:

$$-2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 56.26$$

$$2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 63.74$$

?? (a) We will conduct a 1-sample t-test. H_0: $\mu = 5$. H_A: $\mu < 5$. We'll use $\alpha = 0.05$. This is a random sample, so the observations are independent. To proceed, we assume the distribution of years of piano lessons is approximately normal. $SE = 2.2/\sqrt{30} = 0.402$. The test statistic is $T = (4.6 - 5)/SE = -0.995$. $df = 30 - 1 = 29$. The p-value is about 0.164, which is bigger than $\alpha = 0.05$ and we do not reject H_0. That is, we do not have sufficiently strong evidence to reject Georgianna's claim that the average is (at least) 5 years.
(b) Using $SE = 0.402$ and $t^\star_{df=29} = 2.045$, the confidence interval is (3.78, 5.42). We are 95% confident that the average number of years a child takes piano lessons in this city is between 3.78 and 5.42 years.
(c) They agree, since we did not reject the null hypothesis and the null value of 5 was in the t-interval.

?? Assuming the population standard deviation is known, the margin of error will be $1.96 \times 100/\sqrt{n}$. We want this value to be less than 10, which leads to $n \geq 384.16$, meaning we need a sample size of at least 385 (round up for sample size calculations!).

?? Paired, data are recorded in the same cities at two different time points. The temperature in a city at one point is not independent of the temperature in the same city at another time point.

?? (a) Since it's the same students at the beginning and the end of the semester, there is a pairing between the datasets, for a given student their beginning and end of semester grades are dependent. (b) Since the subjects were sampled randomly, each observation in the men's group does not have a special correspondence with exactly one observation in the other (women's) group. (c) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester artery thickness are dependent. (d) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester weights are dependent.

?? (a) For each observation in one data set, there is exactly one specially corresponding observation in the other data set for the same geographic location. The data are paired. Let $diff = 2018 - 1948$. (b) $H_0 : \mu_{\text{diff}} = 0$ (On average, the number of days exceeding 90°F in 1948 and 2018 for NOAA stations was the same.) $H_A : \mu_{\text{diff}} > 0$ (On average, there were more days exceeding 90°F in 2018 than 1948 for NOAA stations.) (c) Locations were randomly sampled, so independence is reasonable. The sample size is at least 30, so we're just looking for particularly extreme outliers: none are present (the observation off left in the histogram would be considered a clear outlier, but not a particularly extreme one). Therefore, the conditions are satisfied. (d) $SE = 17.2/\sqrt{197} = 1.23$. $T = \frac{2.9-0}{1.23} = 2.36$ with degrees of freedom $df = 197 - 1 = 196$. This leads to a p-value of about 0.019. (e) Since the p-value is less than 0.05, we reject H_0. The data provide strong evidence that NOAA stations observed more 90°F days in 2018 than in 1948. (f) Type 1 Error, since we may have incorrectly rejected H_0. This error would mean that NOAA stations did not actually observe a decrease, but the sample we took just so happened to make it appear that this was the case.

?? (a) $SE = 1.23$ and $z^\star = 1.65$. $2.9 \pm 1.65 \times 1.23 \rightarrow (0.87, 4.93)$.
(b) We are 90% confident that there was an increase of 0.87 to 4.93 in the average number of days that hit 90°F in 2018 relative to 1948 for NOAA stations.
(c) Yes, since the interval lies entirely above 0.

?? (a) These data are paired. For example, the Friday the 13th in say, September 1991, would probably be more similar to the Friday the 6th in September 1991 than to Friday the 6th in another month or year.
(b) Let $\mu_{diff} = \mu_{sixth} - \mu_{thirteenth}$. $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$.
(c) Independence: The months selected are not random. However, if we think these dates are roughly equivalent to a simple random sample of all such Friday 6th/13th date pairs, then independence is reasonable. To proceed, we must make this strong assumption, though we should note this assumption in any reported results. Normality: With fewer than 10 observations, we would need to see clear outliers to be concerned. There is a borderline outlier on the right of the histogram of the differences, so we would want to report this in formal analysis results.
(d) $T = 4.94$ for $df = 10 - 1 = 9 \rightarrow$ p-value $= 0.001$.
(e) Since p-value < 0.05, reject H_0. The data provide strong evidence that the average number of cars at the intersection is higher on Friday the 6^{th} than on Friday the 13^{th}. (We should exercise caution about generalizing the interpretation to all intersections or roads.)
(f) If the average number of cars passing the intersection actually was the same on Friday the 6^{th} and 13^{th}, then the probability that we would observe a test statistic so far from zero is less than 0.01.
(g) We might have made a Type 1 Error, i.e. incorrectly rejected the null hypothesis.

?? (a) $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$. $T = -2.71$. $df = 5$. p-value $= 0.042$. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average number of traffic accident related emergency room admissions are different between Friday the 6^{th} and Friday the 13^{th}. Furthermore, the data indicate that the direction of that difference is that accidents are lower on Friday the 6^{th} relative to Friday the 13^{th}.
(b) (-6.49, -0.17).
(c) This is an observational study, not an experiment, so we cannot so easily infer a causal intervention implied by this statement. It is true that there is a difference. However, for example, this does not mean that a responsible adult going out on Friday the 13^{th} has a higher chance of harm than on any other night.

?? (a) Chicken fed linseed weighed an average of 218.75 grams while those fed horsebean weighed an average of 160.20 grams. Both distributions are relatively symmetric with no apparent outliers. There is more variability in the weights of chicken fed linseed.
(b) $H_0 : \mu_{ls} = \mu_{hb}$. $H_A : \mu_{ls} \neq \mu_{hb}$.
We leave the conditions to you to consider.
$T = 3.02$, $df = min(11, 9) = 9 \rightarrow$ p-value $= 0.014$. Since p-value < 0.05, reject H_0. The data provide strong evidence that there is a significant difference between the average weights of chickens that were fed linseed and horsebean.
(c) Type 1 Error, since we rejected H_0.
(d) Yes, since p-value > 0.01, we would not have rejected H_0.

?? $H_0 : \mu_C = \mu_S$. $H_A : \mu_C \neq \mu_S$. $T = 3.27$, $df = 11 \rightarrow$ p-value $= 0.007$. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean (with weights from casein being higher). Since this is a randomized experiment, the observed difference can be attributed to the diet.

?? Let $\mu_{diff} = \mu_{pre} - \mu_{post}$. $H_0 : \mu_{diff} = 0$: Treatment has no effect. $H_A : \mu_{diff} \neq 0$: Treatment has an effect on P.D.T. scores, either positive or negative. Conditions: The subjects are randomly assigned to treatments, so independence within and between groups is satisfied. All three sample sizes are smaller than 30, so we look for clear outliers. There is a borderline outlier in the first treatment group. Since it is borderline, we will proceed, but we should report this caveat with any results. For all three groups: $df = 13$. $T_1 = 1.89 \rightarrow$ p-value $= 0.081$, $T_2 = 1.35 \rightarrow$ p-value $= 0.200$), $T_3 = -1.40 \rightarrow$ (p-value $= 0.185$). We do not reject the null hypothesis for any of these groups. As earlier noted, there is some uncertainty about if the method applied is reasonable for the first group.

?? $H_0 : \mu_T = \mu_C$. $H_A : \mu_T \neq \mu_C$. $T = 2.24$, $df = 21 \rightarrow$ p-value $= 0.036$. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average food consumption by the patients in the treatment and control groups are different. Furthermore, the data indicate patients in the distracted eating (treatment) group consume more food than patients in the control group.

?? False. While it is true that paired analysis requires equal sample sizes, only having the equal sample sizes isn't, on its own, sufficient for doing a paired test. Paired tests require that there be a special correspondence between each pair of observations in the two groups.

?? (a) We are building a distribution of sample statistics, in this case the sample mean. Such a distribution is called a sampling distribution. (b) Because we are dealing with the distribution of sample means, we need to check to see if the Central Limit Theorem applies. Our sample size is greater than 30, and we are told that random sampling is employed. With these conditions met, we expect that the distribution of the sample mean will be nearly normal and therefore symmetric. (c) Because we are dealing with a sampling distribution, we measure its variability with the standard error. $SE = 18.2/\sqrt{45} = 2.713$. (d) The sample means will be more variable with the smaller sample size.

?? Independence: it is a random sample, so we can assume that the students in this sample are independent of each other with respect to number of exclusive relationships they have been in. Notice that there are no students who have had no exclusive relationships in the sample, which suggests some student responses are likely missing (perhaps only positive values were reported). The sample size is at least 30, and there are no particularly extreme outliers, so the normality condition is reasonable. 90% CI: (2.97, 3.43). We are 90% confident that undergraduate students have been in 2.97 to 3.43 exclusive relationships, on average.

?? The hypotheses should be about the population mean (μ), not the sample mean. The null hypothesis should have an equal sign and the alternative hypothesis should be about the null hypothesized value, not the observed sample mean. Correction:

$$H_0 : \mu = 10 \; hours$$
$$H_A : \mu \neq 10 \; hours$$

Because the change could go either way, we use a two-sided H_A.

8 Introduction to linear regression

8.1 (a) The residual plot will show randomly distributed residuals around 0. The variance is also approximately constant. (b) The residuals will show a fan shape, with higher variability for smaller x. There will also be many points on the right above the line. There is trouble with the model being fit here.

8.3 (a) Strong relationship, but a straight line would not fit the data. (b) Strong relationship, and a linear fit would be reasonable. (c) Weak relationship, and trying a linear fit would be reasonable. (d) Moderate relationship, but a straight line would not fit the data. (e) Strong relationship, and a linear fit would be reasonable. (f) Weak relationship, and trying a linear fit would be reasonable.

8.5 (a) Exam 2 since there is less of a scatter in the plot of final exam grade versus exam 2. Notice that the relationship between Exam 1 and the Final Exam appears to be slightly nonlinear. (b) Exam 2 and the final are relatively close to each other chronologically, or Exam 2 may be cumulative so has greater similarities in material to the final exam. Answers may vary.

8.7 (a) $r = -0.7 \rightarrow$ (4). (b) $r = 0.45 \rightarrow$ (3). (c) $r = 0.06 \rightarrow$ (1). (d) $r = 0.92 \rightarrow$ (2).

8.9 (a) The relationship is positive, weak, and possibly linear. However, there do appear to be some anomalous observations along the left where several students have the same height that is notably far from the cloud of the other points. Additionally, there are many students who appear not to have driven a car, and they are represented by a set of points along the bottom of the scatterplot. (b) There is no obvious explanation why simply being tall should lead a person to drive faster. However, one confounding factor is gender. Males tend to be taller than females on average, and personal experiences (anecdotal) may suggest they drive faster. If we were to follow-up on this suspicion, we would find that sociological studies confirm this suspicion. (c) Males are taller on average and they drive faster. The gender variable is indeed an important confounding variable.

8.11 (a) There is a somewhat weak, positive, possibly linear relationship between the distance traveled and travel time. There is clustering near the lower left corner that we should take special note of. (b) Changing the units will not change the form, direction or strength of the relationship between the two variables. If longer distances measured in miles are associated with longer travel time measured in minutes, longer distances measured in kilometers will be associated with longer travel time measured in hours. (c) Changing units doesn't affect correlation: $r = 0.636$.

8.13 (a) There is a moderate, positive, and linear relationship between shoulder girth and height. (b) Changing the units, even if just for one of the variables, will not change the form, direction or strength of the relationship between the two variables.

8.15 In each part, we can write the woman's age as a linear function of the spouse's age.
(a) $age_W = age_S + 3$.
(b) $age_W = age_S - 2$.
(c) $age_W = 2 \times age_S$.
Since the slopes are positive and these are perfect linear relationships, the correlation will be exactly 1 in all three parts. An alternative way to gain insight into this solution is to create a mock data set, e.g. 5 women aged 26, 27, 28, 29, and 30, then find the spouses ages for each women in each part and create a scatterplot.

8.17 Correlation: no units. Intercept: kg. Slope: kg/cm.

8.19 Over-estimate. Since the residual is calculated as *observed* − *predicted*, a negative residual means that the predicted value is higher than the observed value.

8.21 (a) There is a positive, very strong, linear association between the number of tourists and spending. (b) Explanatory: number of tourists (in thousands). Response: spending (in millions of US dollars). (c) We can predict spending for a given number of tourists using a regression line. This may be useful information for determining how much the country may want to spend in advertising abroad, or to forecast expected revenues from tourism. (d) Even though the relationship appears linear in the scatterplot, the residual plot actually shows a nonlinear relationship. This is not a contradiction: residual plots can show divergences from linearity that can be difficult to see in a scatterplot. A simple linear model is inadequate for modeling these data. It is also important to consider that these data are observed sequentially, which means there may be a hidden structure not evident in the current plots but that is important to consider.

8.23 (a) First calculate the slope: $b_1 = R \times s_y/s_x = 0.636 \times 113/99 = 0.726$. Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$: $\bar{y} = b_0 + b_1 \times \bar{x}$. Plug in $\bar{x}$, $\bar{y}$, and b_1, and solve for b_0: 51. Solution: $\widehat{travel\ time} = 51 + 0.726 \times distance$. (b) b_1: For each additional mile in distance, the model predicts an additional 0.726 minutes in travel time. b_0: When the distance traveled is 0 miles, the travel time is expected to be 51 minutes. It does not make sense to have a travel distance of 0 miles in this context. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself. (c) $R^2 = 0.636^2 = 0.40$. About 40% of the variability in travel time is accounted for by the model, i.e. explained by the distance traveled. (d) $\widehat{travel\ time} = 51 + 0.726 \times distance = 51 + 0.726 \times 103 \approx 126$ minutes. (Note: we should be cautious in our predictions with this model since we have not yet evaluated whether it is a well-fit model.) (e) $e_i = y_i - \hat{y}_i = 168 - 126 = 42$ minutes. A positive residual means that the model underestimates the travel time. (f) No, this calculation would require extrapolation.

8.25 (a) $\widehat{murder} = -29.901 + 2.559 \times poverty\%$. (b) Expected murder rate in metropolitan areas with no poverty is -29. 901 per million. This is obviously not a meaningful value, it just serves to adjust the height of the regression line. (c) For each additional percentage increase in poverty, we expect murders per million to be higher on average by 2.559. (d) Poverty level explains 70.52% of the variability in murder rates in metropolitan areas. (e) $\sqrt{0.7052} = 0.8398$.

8.27 (a) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. It is also an influential point since, without that observation, the regression line would have a very different slope.
(b) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. However, it does not appear to be affecting the line much, so it is not an influential point.
(c) The observation is in the center of the data (in the x-axis direction), so this point does *not* have high leverage. This means the point won't have much effect on the slope of the line and so is not an influential point.

8.29 (a) There is a negative, moderate-to-strong, somewhat linear relationship between percent of families who own their home and the percent of the population living in urban areas in 2010. There is one outlier: a state where 100% of the population is urban. The variability in the percent of homeownership also increases as we move from left to right in the plot.
(b) The outlier is located in the bottom right corner, horizontally far from the center of the other points, so it is a point with high leverage. It is an influential point since excluding this point from the analysis would greatly affect the slope of the regression line.

8.31 (a) The relationship is positive, non-linear, and somewhat strong. Due to the non-linear form of the relationship and the clear non-constant variance in the residuals, a linear model is not appropriate for modeling the relationship between year and price. (b) The logged model is a much better fit: the scatter plot shows a linear relationships and the residuals do not appear to have a pattern. (c) For each year increase in the year of the truck (for each year the truck is newer) we would expect the price of the truck to increase on average by a factor of $e^{0.137} \approx 1.15$, i.e. by 15%.

8.33 (a) The relationship is positive, moderate-to-strong, and linear. There are a few outliers but no points that appear to be influential.
(b) $\widehat{weight} = -105.0113 + 1.0176 \times height$.
Slope: For each additional centimeter in height, the model predicts the average weight to be 1.0176 additional kilograms (about 2.2 pounds).
Intercept: People who are 0 centimeters tall are expected to weigh - 105.0113 kilograms. This is obviously not possible. Here, the y- intercept serves only to adjust the height of the line and is meaningless by itself.
(c) H_0: The true slope coefficient of height is zero ($\beta = 0$).
H_A: The true slope coefficient of height is different than zero ($\beta \neq 0$).
The p-value for the two-sided alternative hypothesis ($\beta \neq 0$) is incredibly small, so we reject H_0. The data provide convincing evidence that height and weight are positively correlated. The true slope parameter is indeed greater than 0.
(d) $R^2 = 0.72^2 = 0.52$. Approximately 52% of the variability in weight can be explained by the height of individuals.

8.35 (a) H_0: $\beta = 0$. H_A: $\beta \neq 0$. The p-value, as reported in the table, is incredibly small and is smaller than 0.05, so we reject H_0. The data provide convincing evidence that women's and spouses' heights are positively correlated.
(b) $\widehat{height}_S = 43.5755 + 0.2863 \times height_W$.
(c) Slope: For each additional inch in woman's height, the spouse's height is expected to be an additional 0.2863 inches, on average. Intercept: Women who are 0 inches tall are predicted to have spouses who are 43.5755 inches tall. The intercept here is meaningless, and it serves only to adjust the height of the line.
(d) The slope is positive, so r must also be positive. $r = \sqrt{0.09} = 0.30$.
(e) 63.2612. Since R^2 is low, the prediction based on this regression model is not very reliable.
(f) No, we should avoid extrapolating.

8.37 (a) $H_0 : \beta = 0; H_A : \beta \neq 0$ (b) The p-value for this test is approximately 0, therefore we reject H_0. The data provide convincing evidence that poverty percentage is a significant predictor of murder rate. (c) $n = 20, df = 18, T^*_{18} = 2.10$; $2.559 \pm 2.10 \times 0.390 = (1.74, 3.378)$; For each percentage point poverty is higher, murder rate is expected to be higher on average by 1.74 to 3.378 per million.
(d) Yes, we rejected H_0 and the confidence interval does not include 0.

8.39 (a) True. (b) False, correlation is a measure of the linear association between any two numerical variables.

8.41 There is an upwards trend. However, the variability is higher for higher calorie counts, and it looks like there might be two clusters of observations above and below the line on the right, so we should be cautious about fitting a linear model to these data.

8.43 (a) $r = -0.72 \rightarrow$ (2) (b) $r = 0.07 \rightarrow$ (4) (c) $r = 0.86 \rightarrow$ (1) (d) $r = 0.99 \rightarrow$ (3)

8.45 (a) There is a weak-to-moderate, positive, linear association between height and volume. There also appears to be some non-constant variance since the volume of trees is more variable for taller trees. (b) There is a very strong, positive association between diameter and volume. The relationship may include slight curvature. (c) Since the relationship is stronger between volume and diameter, using diameter would be preferred. However, as mentioned in part (b), the relationship between volume and diameter may not be, and so we may benefit from a model that properly accounts for nonlinearity.

Appendix B

Data sets within the text

Each data set within the text is described in this appendix, and there is a corresponding page for each of these data sets at **openintro.org/data**. This page also includes additional data sets that can be used for honing your skills. Each data set has its own page with the following information:

- Description of each data set.
- Detailed overview of each data set's variables.
- CSV download.
- R object file download.

Over time we will also expand the information available on these pages.

Chapter 1: ??

?? `stent30`, `stent365` → The stent data is split across two data sets, one for the 0-30 day and one for the 0-365 day results.
Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. www.nejm.org/doi/full/10.1056/NEJMoa1105335.
NY Times article: www.nytimes.com/2011/09/08/health/research/08stent.html.

?? `loan50`, `loans_full_schema` → This data comes from Lending Club (lendingclub.com), which provides a large set of data on the people who received loans through their platform. The data used in the textbook comes from a sample of the loans made in Q1 (Jan, Feb, March) 2018.

?? `county`, `county_complete` → These data come from several government sources. For those variables included in the county data set, only the most recent data is reported, as of what was available in late 2018. Data prior to 2011 is all from census.gov, where the specific Quick Facts page providing the data is no longer available. The more recent data comes from USDA (ers.usda.gov), Bureau of Labor Statistics (bls.gov/lau), SAIPE (census.gov/did/www/saipe), and American Community Survey (census.gov/programs-surveys/acs).

?? The Nurses' Health Study was mentioned. For more information on this data set, see www.channing.harvard.edu/nhs

?? The study we had in mind when discussing the simple randomization (no blocking) study was Anturane Reinfarction Trial Research Group. 1980. *Sulfinpyrazone in the prevention of sudden death after myocardial infarction.* New England Journal of Medicine 302(5):250-256.

Chapter 2: Summarizing data

2.1 `loan50`, `county` → These data sets are described in the data for Chapter **??**.

2.4 `loan50`, `county` → These data sets are described in the data for Chapter **??**.

2.5 `malaria` → Lyke et al. 2017. PfSPZ vaccine induces strain-transcending T cells and durable protection against heterologous controlled human malaria infection. PNAS 114(10):2711-2716. www.pnas.org/content/114/10/2711

Chapter 3: Probability and probability distributions

3.1 `loan50`, `county` → These data sets are described in the data for Chapter **??**.

3.1 `playing_cards` → A table describing the 52 cards in a standard deck.

3.2 `family_college` → A simulated data set based on real population summaries at nces.ed.gov/pubs2001/2001126.pdf.

3.2 `smallpox` → Fenner F. 1988. Smallpox and Its Eradication (History of International Public Health, No. 6). Geneva: World Health Organization. ISBN 92-4-156110-6.

3.2 Mammogram screening, probabilities. → The probabilities reported were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

3.4 `stocks_18` → Monthly returns for Caterpillar, Exxon Mobil Corp, and Google for November 2015 to October 2018.

?? `fcid` → This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.

Chapter 4: ??

2.3 SAT and ACT score distributions → The SAT score data comes from the 2018 distribution, which is provided at
reports.collegeboard.org/pdf/2018-total-group-sat-suite-assessments-annual-report.pdf
The ACT score data is available at
act.org/content/dam/act/unsecured/documents/cccr2018/P_99_999999_N_S_N00_ACT-GCPR_National.pdf
We also acknowledge that the actual ACT score distribution is *not* nearly normal. However, since the topic is very accessible, we decided to keep the context and examples.

2.3 `possum` → The distribution parameters are based on a sample of possums from Australia and New Guinea. The original source of this data is as follows. Lindenmayer DB, et al. 1995. *Morphological variation among columns of the mountain brushtail possum, Trichosurus caninus Ogilby (Phalangeridae: Marsupiala)*. Australian Journal of Zoology 43: 449-458.

2.3 `male_heights_fcid` → This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.

2.3 `nba_players_19` → Summary information from the NBA players for the 2018-2019 season. Data were retrieved from www.nba.com/players.

2.3 `poker` → Poker winnings (and losses) for 50 days by a professional poker player, which represents their first 50 days trying to play for a living. Anonymity has been requested by the player.

?? `run17`, `run17samp` → www.cherryblossom.org

Chapter 5: ??

?? `pew_energy_2018` → The actual data has more observations than were referenced in this chapter. That is, we used a subsample since it helped smooth some of the examples to have a bit more variability. The `pew_energy_2018` data set represents the full data set for each of the different energy source questions, which covers solar, wind, offshore drilling, hydrolic fracturing, and nuclear energy. The statistics used to construct the data are from the following page:

www.pewinternet.org/2018/05/14/majorities-see-government-efforts-to-protect-the-environment-as-insufficient/

?? `pew_energy_2018` → See the details for this data set above in the Section **??** data section.

?? `ebola_survey` → In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon thereafter, an NBC 4 New York/The Wall Street Journal/Marist Poll found that 82% of New Yorkers favored a "mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient". This poll included responses of 1,042 New York adults between Oct 26th and 28th, 2014. Poll ID NY141026 on maristpoll.marist.edu.

?? `pew_energy_2018` → See the details for this data set above in the Section **??** data section.

Chapter 6: ??

?? Nuclear energy → A Gallup poll of 1,019 adults in the US, conducted in March of 2016, found that 54% of respondents oppose nuclear energy. This was the first time since Gallup first asked the question in 1994 that a majority of respondents said they oppose nuclear energy.
https://news.gallup.com/poll/190064/first-time-majority-oppose-nuclear-energy.aspx

?? Supreme Court → The Gallup organization began measuring the public's view of the Supreme Court's job performance in 2000, and has measured it every year since then with the question: "Do you approve or disapprove of the way the Supreme Court is handling its job?". In 2018, the Gallup poll randomly sampled 1,033 adults in the U.S. and found that 53% of them approved.
https://news.gallup.com/poll/237269/supreme-court-approval-highest-2009.aspx

?? Life on other planets → A February 2018 Marist Poll reported: "Many Americans (68%) think there is intelligent life on other planets". The results were based on a random sample of 1,033 adults in the U.S.
http://maristpoll.marist.edu/212-are-americans-poised-for-an-alien-invasion

?? `cpr` → Böttiger et al. *Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial.* The Lancet, 2001.

?? `fish_oil_18` → Manson JE, et al. 2018. Marine n-3 Fatty Acids and Prevention of Cardiovascular Disease and Cancer. NEJMoa1811403.

?? `mammogram` → Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial.* BMJ 2014;348:g366.

?? `drone_blades` → The quality control data set for quadcopter drone blades is a made-up data set for an example. We provide the simulated data in the `drone_blades` data set.

?? M&Ms → Starting at the end of 2016, Rick Wicklin, a statistician working at the statistical software company SAS, collected a sample of 712 candies, or about 1.5 pounds, and counted how many there were of each color.
https://qz.com/918008/the-color-distribution-of-mms-as-determined-by-a-phd-in-statistics

?? `ask` → Minson JA, Ruedy NE, Schweitzer ME. *There is such a thing as a stupid question: Question disclosure in strategic communication.*
opim.wharton.upenn.edu/DPlab/papers/workingPapers/
Minson_working_Ask%20(the%20Right%20Way)%20and%20You%20Shall%20Receive.pdf

?? `diabetes2` → Zeitler P, et al. 2012. A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes. N Engl J Med.

Chapter 7: ??

?? Risso's dolphins → Endo T and Haraguchi K. 2009. High mercury levels in hair samples from residents of Taiji, a Japanese whaling town. Marine Pollution Bulletin 60(5):743-747.

Taiji was featured in the movie *The Cove*, and it is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we will assume these 19 dolphins represent a simple random sample from those dolphins.

?? Croaker white fish → www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

?? `run17`, `run17samp` → www.cherryblossom.org

?? `textbooks`, `ucla_textbooks_f18` → Data were collected by OpenIntro staff in 2010 and again in 2018. For the 2018 sample, we sampled 201 UCLA courses. Of those, 68 required books that could be found on Amazon. The websites where information was retrieved: sa.ucla.edu/ro/public/soc, ucla.verbacompare.com, and amazon.com.

?? Jennifer-John → Bertrand M, Mullainathan S. 2004. *Science faculty's subtle gender biases favor male students.* PNAS October 9, 2012 109 (41) 16474-16479. https://www.pnas.org/content/109/41/16474

?? `resume` → Bertrand M, Mullainathan S. 2004. *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.* The American Economic Review 94:4 (991-1013). www.nber.org/papers/w9873

?? `stem_cells` → Menard C, et al. 2005. Transplantation of cardiac-committed mouse embryonic stem cells to infarcted sheep myocardium: a preclinical study. The Lancet: 366:9490, p1005-1012. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(05)67380-1/fulltext

Chapter 8: Introduction to linear regression

8.1 `simulated_scatter` → Fake data used for the first three plots. The perfect linear plot uses group 4 data, where `group` variable in the data set (Figure 8.1). The group of 3 imperfect linear plots use groups 1-3 (Figure 8.2). The sinusoidal curve uses group 5 data (Figure 8.3). The group of 3 scatterplots with residual plots use groups 6-8 (Figure 8.8). The correlation plots uses groups 9-19 data (Figures 8.9 and 8.10).

8.1 `possum` → This data is described in the data for Chapter **??**.

8.2 `elmhurst` → These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled *What Students Really Pay to Go to College* published online by *The Chronicle of Higher Education*: chronicle.com/article/What-Students-Really-Pay-to-Go/131435.

8.2 `simulated_scatter` → The plots for things that can go wrong uses groups 20-23 (Figure 8.26).

8.2 `mariokart` → Auction data from Ebay (ebay.com) for the game Mario Kart for the Nintendo Wii. This data set was collected in early October, 2009.

8.2 `simulated_scatter` → The plots for types of outliers uses groups 24-29 (Figure 8.19).

8.4 `midterms_house` → Data was retrieved from Wikipedia.

8.3 `county`, `county_complete` → This data is described in the data for Chapter **??**.

Appendix C

Distribution tables

C.1 Random Number Table

	Column							
Row	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
1	44394	76100	85973	26853	07080	91603	00476	19681
2	61578	75037	54792	74216	31952	31235	31258	57886
3	18529	73285	95291	49606	67174	95905	33679	75811
4	81238	18321	71085	08284	39318	31434	26173	07440
5	11173	58878	25516	15058	48639	52723	95864	89673
6	96737	95194	14419	22202	92867	73525	94382	29927
7	63514	55066	65162	96016	91723	21160	24285	33264
8	35087	57036	10001	39424	50536	77380	45042	48180
9	00148	73933	49369	32403	53850	16291	93619	27557
10	28999	76232	32637	95697	63679	54506	11299	94294
11	37911	50834	10927	74075	26558	42311	36483	71820
12	33624	82379	03625	58336	27390	00586	06344	89625
13	93282	63059	10830	89432	26917	31555	51793	18718
14	57429	71933	80329	56521	97594	92651	14819	86546
15	65029	24328	06826	61448	54760	09351	73930	99564
16	14779	23173	97183	59835	69580	94653	55095	80666
17	52072	12187	35360	82925	44923	44532	18251	96991
18	76282	91849	17138	59554	35476	67007	02484	10122
19	46561	33015	04577	02178	32915	35912	48974	92985
20	70623	36097	48780	06921	60683	22461	36175	61281
21	03605	08541	17546	85790	48413	69382	89785	80206
22	46147	07603	92057	87609	52670	96255	96660	83167
23	09547	77804	95099	22158	53279	23161	72675	92804
24	12899	05005	86667	72331	09114	28187	97404	26750
25	21223	38353	56970	48965	58371	02697	61417	54746
26	35770	35697	32281	53514	10854	16778	56447	46965
27	04243	65817	81819	64381	83509	44316	56316	47742
28	56989	05587	79995	36598	02316	81627	50104	47720
29	53233	48698	59304	63566	25352	03322	29938	82306
30	20232	30909	77126	50041	96500	24033	77422	20150

C.2 Normal Probability Table

A **normal probability table** may be used to find percentiles of a normal distribution using a Z-score, or vice-versa. Such a table lists Z-scores and the corresponding percentiles. An abbreviated probability table is provided in Figure C.1 that we'll use for the examples in this appendix. A full table may be found on page 266.

	Second decimal place of Z									
Z	*0.00*	0.01	0.02	0.03	**0.04**	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	**0.7995**	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	*0.8413*	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure C.1: A section of the normal probability table. The percentile for a normal random variable with $Z = 1.00$ has been *highlighted*, and the percentile closest to 0.8000 has also been **highlighted**.

When using a normal probability table to find a percentile for Z (rounded to two decimals), identify the proper row in the normal probability table up through the first decimal, and then determine the column representing the second decimal value. The intersection of this row and column is the percentile of the observation. For instance, the percentile of $Z = 0.45$ is shown in row 0.4 and column 0.05 in Figure C.1: 0.6736, or the 67.36^{th} percentile.

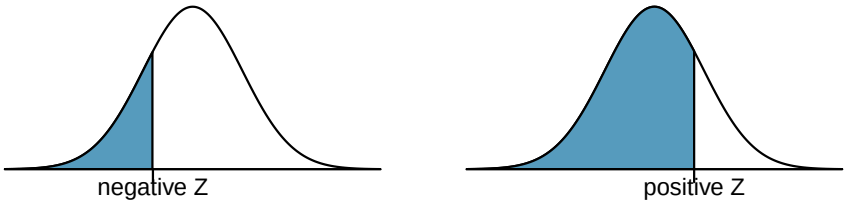
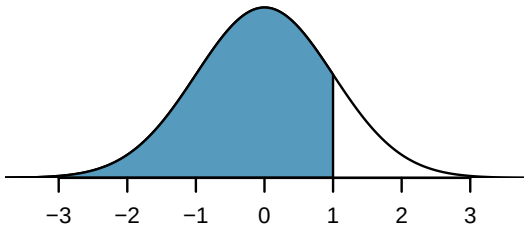


Figure C.2: The area to the left of Z represents the percentile of the observation.

EXAMPLE C.1

SAT scores follow a normal distribution, $N(1100, 200)$. Ann earned a score of 1300 on her SAT with a corresponding Z-score of $Z = 1$. She would like to know what percentile she falls in among all SAT test-takers.

Ann's **percentile** is the percentage of people who earned a lower SAT score than her. We shade the area representing those individuals in the following graph:



The total area under the normal curve is always equal to 1, and the proportion of people who scored below Ann on the SAT is equal to the *area* shaded in the graph. We find this area by looking in row 1.0 and column 0.00 in the normal probability table: 0.8413. In other words, Ann is in the 84^{th} percentile of SAT takers.

EXAMPLE C.2

How do we find an upper tail area?

The normal probability table *always* gives the area to the left. This means that if we want the area to the right, we first find the lower tail and then subtract it from 1. For instance, 84.13% of SAT takers scored below Ann, which means 15.87% of test takers scored higher than Ann.

We can also find the Z-score associated with a percentile. For example, to identify Z for the 80^{th} percentile, we look for the value closest to 0.8000 in the middle portion of the table: 0.7995. We determine the Z-score for the 80^{th} percentile by combining the row and column Z values: 0.84.

EXAMPLE C.3

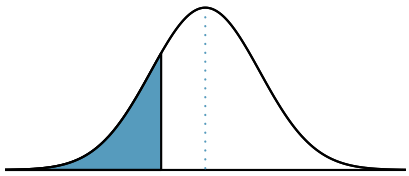
Find the SAT score for the 80^{th} percentile.

We look for the are to the value in the table closest to 0.8000. The closest value is 0.7995, which corresponds to $Z = 0.84$, where 0.8 comes from the row value and 0.04 comes from the column value. Next, we set up the equation for the Z-score and the unknown value x as follows, and then we solve for x:

$$Z = 0.84 = \frac{x - 1100}{200} \quad \rightarrow \quad x = 1268$$

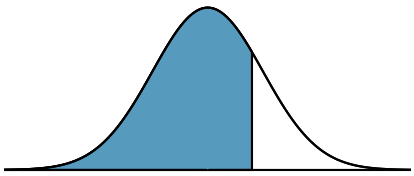
The College Board scales scores to increments of 10, so the 80^{th} percentile is 1270. (Reporting 1268 would have been perfectly okay for our purposes.)

For additional details about working with the normal distribution and the normal probability table, see Section 2.3, which starts on page 50.



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	−1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	−1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	−1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	−1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	−1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	−0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	−0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	−0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	−0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	−0.5
0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	−0.4
0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	−0.3
0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	−0.2
0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	−0.1
0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	−0.0

*For $Z \leq -3.50$, the probability is less than or equal to 0.0002.



Positive Z

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

*For $Z \geq 3.50$, the probability is greater than or equal to 0.9998.

C.3 *t*-Probability Table

A ***t*-probability table** may be used to find tail areas of a t-distribution using a T-score, or vice-versa. Such a table lists T-scores and the corresponding percentiles. A partial ***t*-table** is shown in Figure C.3, and the complete table starts on page 270. Each row in the t-table represents a t-distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t-distribution with $df = 18$, we can examine row 18, which is highlighted in Figure C.3. If we want the value in this row that identifies the T-score (cutoff) for an upper tail of 10%, we can look in the column where *one tail* is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t-distributions are symmetric.

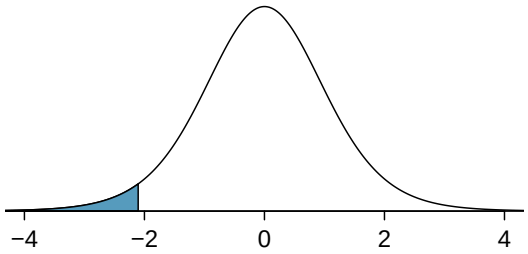
one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	**1.33**	**1.73**	**2.10**	**2.55**	**2.88**
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	⋮	⋮	⋮	⋮	⋮	
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.64	1.96	2.33	2.58

Figure C.3: An abbreviated look at the t-table. Each row represents a different t-distribution. The columns describe the cutoffs for specific tail areas. The row with $df = 18$ has been **highlighted**.

EXAMPLE C.4

What proportion of the t-distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture and shade the area below -2.10:



To find this area, we first identify the appropriate row: $df = 18$. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. That is, 2.5% of the distribution falls below -2.10.

In the next example we encounter a case where the exact T-score is not listed in the table.

EXAMPLE C.5

A t-distribution with 20 degrees of freedom is shown in the left panel of Figure C.4. Estimate the proportion of the distribution falling above 1.65.

(E)

We identify the row in the t-table using the degrees of freedom: $df = 20$. Then we look for 1.65; it is not listed. It falls between the first and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the first and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

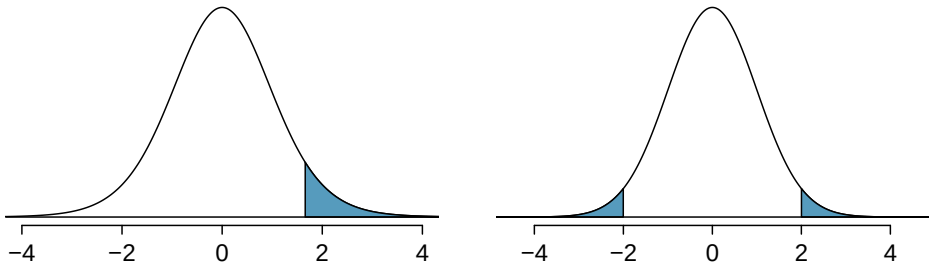


Figure C.4: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 475 degrees of freedom, with the area further than 2 units from 0 shaded.

EXAMPLE C.6

A t-distribution with 475 degrees of freedom is shown in the right panel of Figure C.4. Estimate the proportion of the distribution falling more than 2 units from the mean (above or below).

(E)

As before, first identify the appropriate row: $df = 475$. This row does not exist! When this happens, we use the next smaller row, which in this case is $df = 400$. Next, find the columns that capture 2.00; because $1.97 < 3 < 2.34$, we use the third and fourth columns. Finally, we find bounds for the tail areas by looking at the two tail values: 0.02 and 0.05. We use the two tail values because we are looking for two symmetric tails in the t-distribution.

(G)

GUIDED PRACTICE C.7

What proportion of the t-distribution with 19 degrees of freedom falls above -1.79 units?[1]

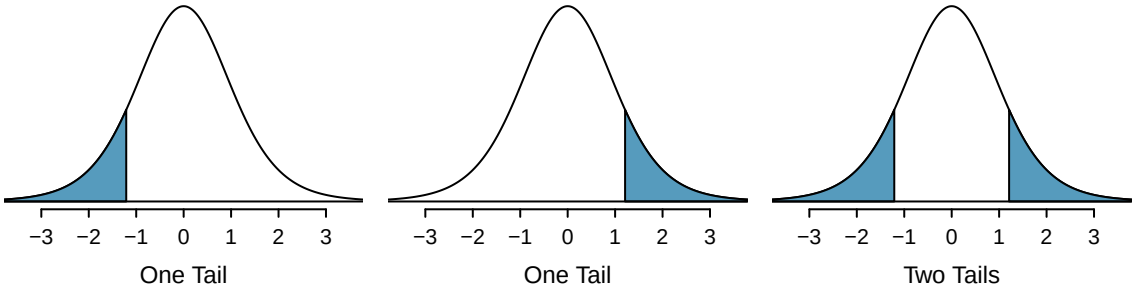
EXAMPLE C.8

Find the value of $t^\star_{18}$ using the t-table, where $t^\star_{18}$ is the cutoff for the t-distribution with 18 degrees of freedom where 95% of the distribution lies between $-t^\star_{18}$ and $+t^\star_{18}$.

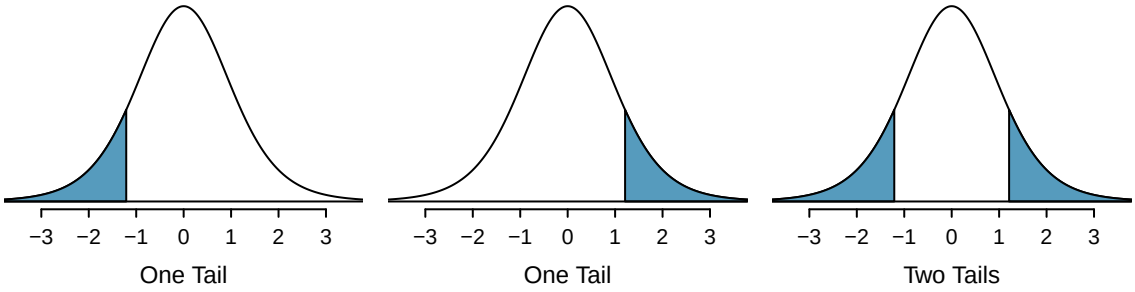
(E)

For a 95% confidence interval, we want to find the cutoff $t^\star_{18}$ such that 95% of the t-distribution is between $-t^\star_{18}$ and $t^\star_{18}$; this is the same as where the two tails have a total area of 0.05. We look in the t-table on page 268, find the column with area totaling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t^\star_{18} = 2.10$.

[1] We find the shaded area *above* -1.79 (we leave the picture to you). The small left tail is between 0.025 and 0.05, so the larger upper region must have an area between 0.95 and 0.975.



one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95
	16	1.34	1.75	2.12	2.58	2.92
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79
	26	1.31	1.71	2.06	2.48	2.78
	27	1.31	1.70	2.05	2.47	2.77
	28	1.31	1.70	2.05	2.47	2.76
	29	1.31	1.70	2.05	2.46	2.76
	30	1.31	1.70	2.04	2.46	2.75



one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	31	1.31	1.70	2.04	2.45	2.74
	32	1.31	1.69	2.04	2.45	2.74
	33	1.31	1.69	2.03	2.44	2.73
	34	1.31	1.69	2.03	2.44	2.73
	35	1.31	1.69	2.03	2.44	2.72
	36	1.31	1.69	2.03	2.43	2.72
	37	1.30	1.69	2.03	2.43	2.72
	38	1.30	1.69	2.02	2.43	2.71
	39	1.30	1.68	2.02	2.43	2.71
	40	1.30	1.68	2.02	2.42	2.70
	41	1.30	1.68	2.02	2.42	2.70
	42	1.30	1.68	2.02	2.42	2.70
	43	1.30	1.68	2.02	2.42	2.70
	44	1.30	1.68	2.02	2.41	2.69
	45	1.30	1.68	2.01	2.41	2.69
	46	1.30	1.68	2.01	2.41	2.69
	47	1.30	1.68	2.01	2.41	2.68
	48	1.30	1.68	2.01	2.41	2.68
	49	1.30	1.68	2.01	2.40	2.68
	50	1.30	1.68	2.01	2.40	2.68
	60	1.30	1.67	2.00	2.39	2.66
	70	1.29	1.67	1.99	2.38	2.65
	80	1.29	1.66	1.99	2.37	2.64
	90	1.29	1.66	1.99	2.37	2.63
	100	1.29	1.66	1.98	2.36	2.63
	150	1.29	1.66	1.98	2.35	2.61
	200	1.29	1.65	1.97	2.35	2.60
	300	1.28	1.65	1.97	2.34	2.59
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.65	1.96	2.33	2.58

C.4 Chi-Square Probability Table

A **chi-square probability table** may be used to find tail areas of a chi-square distribution. The **chi-square table** is partially shown in Figure C.5, and the complete table may be found on page 273. When using a chi-square table, we examine a particular row for distributions with different degrees of freedom, and we identify a range for the area (e.g. 0.025 to 0.05). Note that the chi-square table provides upper tail values, which is different than the normal and t-distribution tables.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	2	2.41	**3.22**	**4.61**	5.99	7.82	9.21	10.60	13.82
	3	*3.66*	*4.64*	6.25	*7.81*	*9.84*	*11.34*	*12.84*	*16.27*
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Figure C.5: A section of the chi-square table. A complete table is in Appendix C.4.

EXAMPLE C.9

Figure C.6(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Figure C.5 to estimate the shaded area.

(E)

This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value – 6.25 – falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure C.6(a) has area 0.1.

This example was unusual, in that we observed the *exact* value in the table. In the next examples, we encounter situations where we cannot precisely estimate the tail area and must instead provide a range of values.

EXAMPLE C.10

Figure C.6(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The area above value 4.3 has been shaded; find this tail area.

(E)

The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure C.6(b) is between 0.1 and 0.2.

EXAMPLE C.11

Figure C.6(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

(E)

Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is *greater than* 0.3.

EXAMPLE C.12

Figure C.6(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.

(E)

The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.

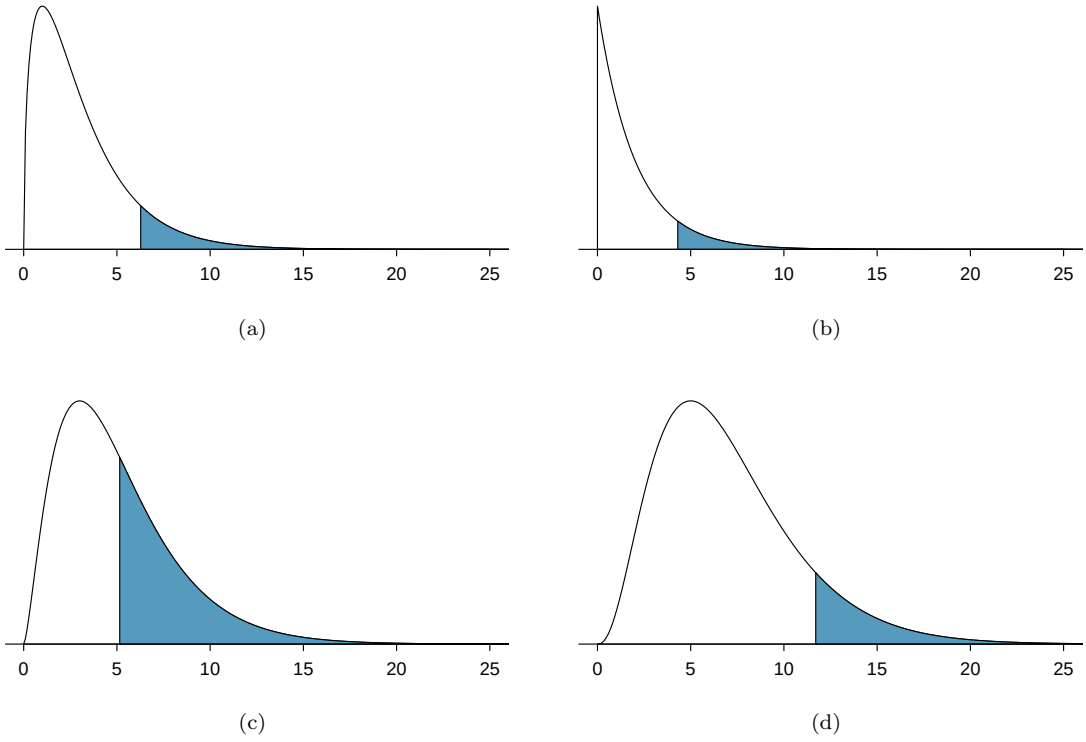


Figure C.6: **(a)** Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. **(b)** 2 degrees of freedom, area above 4.3 shaded. **(c)** 5 degrees of freedom, area above 5.1 shaded. **(d)** 7 degrees of freedom, area above 11.7 shaded.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
	12	14.01	15.81	18.55	21.03	24.05	26.22	28.30	32.91
	13	15.12	16.98	19.81	22.36	25.47	27.69	29.82	34.53
	14	16.22	18.15	21.06	23.68	26.87	29.14	31.32	36.12
	15	17.32	19.31	22.31	25.00	28.26	30.58	32.80	37.70
	16	18.42	20.47	23.54	26.30	29.63	32.00	34.27	39.25
	17	19.51	21.61	24.77	27.59	31.00	33.41	35.72	40.79
	18	20.60	22.76	25.99	28.87	32.35	34.81	37.16	42.31
	19	21.69	23.90	27.20	30.14	33.69	36.19	38.58	43.82
	20	22.77	25.04	28.41	31.41	35.02	37.57	40.00	45.31
	25	28.17	30.68	34.38	37.65	41.57	44.31	46.93	52.62
	30	33.53	36.25	40.26	43.77	47.96	50.89	53.67	59.70
	40	44.16	47.27	51.81	55.76	60.44	63.69	66.77	73.40
	50	54.72	58.16	63.17	67.50	72.61	76.15	79.49	86.66

Appendix D

Calculator reference, Formulas, and Inference guide

D.1 Calculator reference

Instructions for the TI-83/84 and the Casio fx-9750GII, and their associated videos.

Summarizing 1-variable statistics

- Entering data page 32
- Calculating summary statistics. page 33
- Drawing a box plot page 33

Binomial probabilities

- Computing the binomial coefficient page 154
- Computing the binomial formula page 155
- Computing cumulative binomial probabilities page 155

Finding normal probabilities

- Finding area under the normal curve page 58
- Finding a Z-score that corresponds to a percentile page 60

Inference for a single proportion

- 1-proportion Z-interval page **??**
- 1-proportion Z-test page **??**

Inference for a difference of proportions

- 2-proportion Z-interval page **??**
- 2-proportion Z-test page **??**

Chi-square for one-way tables

- Finding area under chi-square curve page **??**
- Chi-square goodness of fit test page **??**

Chi-square for two-way tables

- Entering data in a two-way table page **??**
- Chi-square test of homogeneity and independence page **??**
- Finding the expected counts page **??**

Inference for a single mean

- 1-sample t-interval page **??**
- 1-sample t-test page **??**

Inference for a mean of differences

- Matched pairs t-test page **??**
- Matched pairs t-interval page **??**

Inference for a difference of means

- 2-sample t-interval page **??**
- 2-sample t-test page **??**

The least squares regression line

- Finding the y-intercept, slope, r, and R^2 page 195
- What to do if you get Dim Mismatch page 196

Inference for the slope of a regression line

- t-interval for the slope page 219
- t-test for the slope page 226

D.2 Formulas

Descriptive Statistics

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{\sum x_i}{n}$$

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

$$r = \frac{1}{n-1}\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$\hat{y} = a + bx$$

$$\bar{y} = a + b\bar{x}$$

$$b = r\frac{s_y}{s_x}$$

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

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

$$\mu_X = E(X) = \sum x_i \cdot P(x_i)$$

$$\sigma_X = \sqrt{\sum (x_i - \mu_x)^2 \cdot P(x_i)}$$

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If X has a binomial distribution with parameters n and p, then:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$$

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Inferential Statistics

$$\text{standardized test statistic: } \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$$

$$\text{confidence interval: point estimate } \pm \text{ critical value} \times SE \text{ of estimate}$$

	parameter	point estimate	SE of estimate	
single proportion	p	$\hat{p}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	when H_0: $p = p_0$, use $\sqrt{\frac{p_0(1-p_0)}{n}}$
diff. of proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	when H_0: $p_1 = p_2$, use $\sqrt{\hat{p}_c(1-\hat{p}_c)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
single mean	μ	$\bar{x}$	$\frac{s}{\sqrt{n}}$	
mean of differences	μ_{diff}	$\bar{x}_{diff}$	$\frac{s_{diff}}{\sqrt{n_{diff}}}$	
difference of means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	
slope of reg. line	β	b	$\frac{s}{s_x\sqrt{n-1}}$	

$$\text{Chi-square test statistic} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

INFERENCE GUIDE

CONFIDENCE INTERVALS

Use **confidence intervals** to **estimate** a parameter with a particular **confidence level, C**.

IDENTIFY: Identify the parameter and the confidence level.

CHOOSE: Choose and name the appropriate interval.

CHECK: Check that conditions for the procedure are met.

CALCULATE:

CI: point estimate ± critical value × *SE* of estimate

df = (if applicable)

(_____ , _____)

CONCLUDE:

We are C% confident that the true [parameter] is between _____ and _____ . (Put the parameter in *context.*)

We have evidence that [...], because [...]. OR

We do not have evidence that [...], because [...].

HYPOTHESIS TESTS

Use **hypothesis tests** to **test** H_0 versus H_A at a particular **significance level, α**.

IDENTIFY: Identify the hypotheses and the significance level.

CHOOSE: Choose and name the appropriate test.

CHECK: Check that conditions for the procedure are met.

CALCULATE:

$$\textbf{standardized test statistic} = \frac{\textbf{point estimate} - \textbf{null value}}{\textbf{SE of estimate}}$$

df = (if applicable)

p-value =

CONCLUDE:

p-value < α, so we reject H_0.

We have evidence that [H_A]. (Put H_A in *context.*)

OR

p-value > α, so we do NOT reject H_0.

We do NOT have evidence that [H_A]. (Put H_A in *context.*)

When the parameter is: **a single proportion *p***

CHOOSE: **1-Proportion Z-Interval** to estimate p, or **1-Proportion Z-Test** to test $H_0: p = p_0$.

CHECK:

- Data come from a random sample or process.
- for CI: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.
  for Test: $np_0 \geq 10$ and $n(1-p_0) \geq 10$.

CALCULATE: (`1-PropZInt` or `1-PropZTest`)

point estimate: sample proportion $\hat{p}$

***SE* of estimate:** for CI, use $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$; for Test, use $\sqrt{\frac{p_0(1-p_0)}{n}}$

When the parameter is: **a single mean μ**

CHOOSE: **1-Sample T-Interval** to estimate μ, or **1-Sample T-Test** to test $H_0: \mu = \mu_0$.

CHECK:

- Data come from a random sample or process.
- $n \geq 30$, OR population known to be nearly normal, OR population could to be nearly normal because data has no excessive skew or outliers (draw graph).

CALCULATE: (`TInterval` or `T-Test`)

point estimate: sample mean $\bar{x}$

***SE* of estimate**: $\frac{s}{\sqrt{n}}$

$df = n - 1$

When the parameter is: **a difference of proportions p_1-p_2**

CHOOSE: **2-Proportion Z-Interval** to estimate $p_1 - p_2$, or **2-Proportion Z-Test** to test $H_0: p_1 = p_2$.

CHECK:

- Data come from 2 independent random samples or 2 randomly assigned treatments.
- $n_1\hat{p}_1 \geq 10,\ n_1(1-\hat{p}_1) \geq 10,$
  $n_2\hat{p}_2 \geq 10,\ n_2(1-\hat{p}_2) \geq 10.$
  Note: use $\hat{p}_c$, the pooled proportion, in place of $\hat{p}_1$ and $\hat{p}_2$ when checking condition for the 2-Proportion Z-Test

CALCULATE: (`2-PropZInt` or `2-PropZTest`)

point estimate: difference of sample proportions $\hat{p}_1 - \hat{p}_2$

***SE* of estimate:**

CI, use $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$; Test, use $\sqrt{\hat{p}_c(1-\hat{p}_c)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

When the parameter is: **a difference of means μ_1-μ_2**

CHOOSE: **2-Sample T-Interval** to estimate $\mu_1 - \mu_2$, or **2-Sample T-Test** to test $H_0: \mu_1 = \mu_2$.

CHECK:

- Data come from 2 independent random samples or 2 randomly assigned treatments.
- $n_1 \geq 30$ and $n_2 \geq 30$, OR *both* populations known to be nearly normal, OR *both* populations could be nearly normal because both data sets have no excessive skew or outliers (draw 2 graphs).

CALCULATE: (`2-SampTInt` or `2-SampTTest`)

point estimate: difference of sample means $\bar{x}_1 - \bar{x}_2$

***SE* of estimate**: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

df: use technology

When the parameter is: **a mean of differences μ_{diff}**

CHOOSE: **Matched Pairs T-Interval** to estimate μ_{diff}, or
Matched Pairs T-Test to test H_0: $\mu_{diff} = 0$.

CHECK:
- There is paired data from a random sample or matched pairs experiment.
- $n_{diff} \geq 30$, OR population of differences known to be nearly normal, OR population of differences could be nearly normal because observed differences have no excessive skew or outliers (draw graph of *differences*).

CALCULATE: (`TInterval` or `T-Test`)
point estimate: mean of sample difference $\bar{x}_{diff}$
***SE* of estimate**: $\frac{s_{diff}}{\sqrt{n_{diff}}}$
$df = n_{diff} - 1$

When the parameter is: **the slope β of a regression line**

CHOOSE: **T-Interval for the slope** to estimate β, or
T-Test for the slope to test H_0: $\beta = 0$.

CHECK:
- There is (*x*, *y*) data from a random sample or experiment.
- The residual plot shows no pattern making a linear model reasonable. (More specifically, the residuals should be independent, nearly normal, and have constant standard deviation.)

CALCULATE: (`LinRegTInt` or `LinRegTTest`)
point estimate: sample slope b
***SE* of estimate**: SE of slope (from computer output)
$df = n - 2$

The χ2 tests for categorical variables: **chi-square statistic** $= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

When comparing the distribution of **one categorical variable to a fixed/specified population distribution**

CHOOSE: **χ2 Goodness of Fit Test**

CHECK:
- Data come from a random sample or process.
- All expected counts ≥ 5. (To calculate expected counts for each category, multiply the sample size by the expected proportion under H_0.)

CALCULATE: (`χ2GOF-Test`)
$\chi^2 =$
$df =$ # of categories – 1

When comparing the distribution of **a categorical variable across 2 or more populations/treatments**

CHOOSE: **χ2 Test for Homogeneity**

CHECK:
- Data come from 2 or more independent random samples or 2 or more randomly assigned treatments.
- All expected counts ≥ 5. (Calculate expected counts and verify this to be true.)

CALCULATE: (`χ2-Test`, then `2ND MATRIX,EDIT,2:[B]` to find expected counts)
$\chi^2 =$
$df =$ (# of rows – 1)(# of cols – 1)

When looking for **association or dependence between two categorical variables**

CHOOSE: **χ2 Test for Independence**

CHECK:
- Data come from a random sample or process.
- All expected counts ≥ 5. (Calculate expected counts and verify this to be true.)

CALCULATE: (`χ2-Test`, then `2ND MATRIX,EDIT,2:[B]` to find expected counts)
$\chi^2 =$
$df =$ (# of rows – 1)(# of cols – 1)