

Chapter 2

Summarizing data

2.1 Examining numerical data

2.2 Considering categorical data

2.3 Case study: malaria vaccine

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



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

2.1 Examining numerical data

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

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

2.1.1 Scatterplots for paired data

A **scatterplot** provides a case-by-case view of data for two numerical variables. In Figure 1.8 on page 16, a scatterplot was used to examine the homeownership rate against the fraction of housing units that were part of multi-unit properties (e.g. apartments) in the `county` data set. Another scatterplot is shown in Figure 2.1, comparing the total income of a borrower (`total_income`) and the amount they borrowed (`loan_amount`) for the `loan50` data set. In any scatterplot, each point represents a single case. Since there are 50 cases in `loan50`, there are 50 points in Figure 2.1.

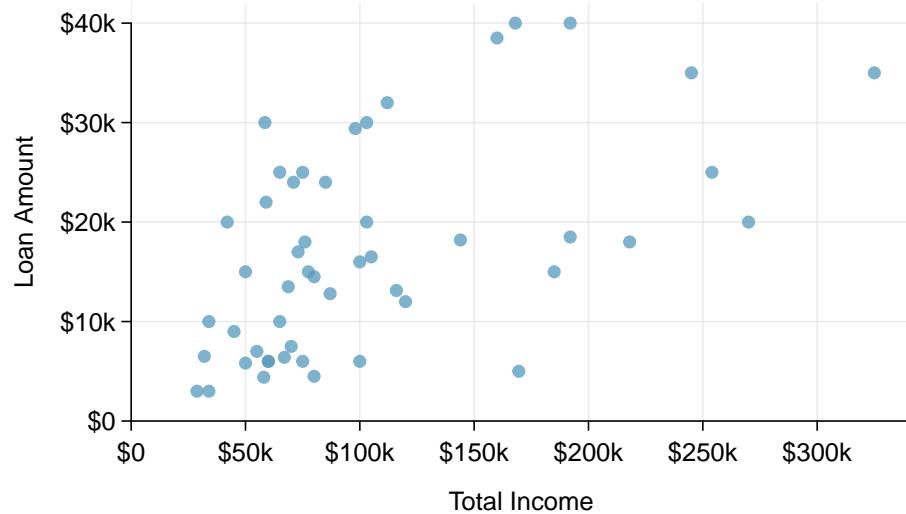


Figure 2.1: A scatterplot of `total_income` versus `loan_amount` for the `loan50` data set.

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

EXAMPLE 2.1

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

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.

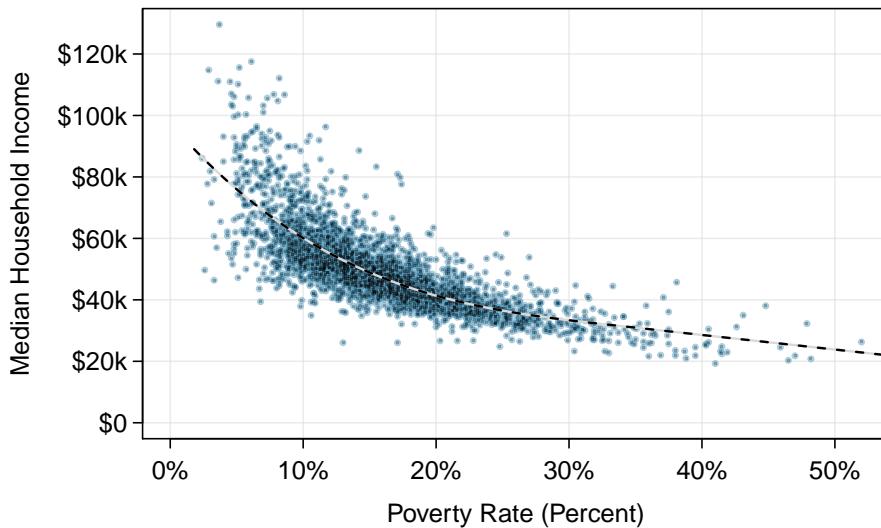


Figure 2.2: A scatterplot of the median household income against the poverty rate for the county data set. A statistical model has also been fit to the data and is shown as a dashed line.



GUIDED PRACTICE 2.2

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



GUIDED PRACTICE 2.3

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

2.1.2 Dot plots and the mean

Sometimes two variables are one too many: only one variable may be of interest. In these cases, a dot plot provides the most basic of displays. A **dot plot** is a one-variable scatterplot; an example using the interest rate of 50 loans is shown in Figure 2.3. A stacked version of this dot plot is shown in Figure 2.4.

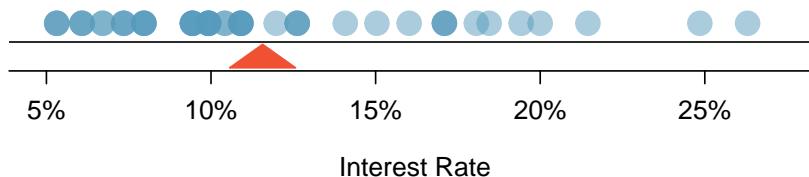


Figure 2.3: A dot plot of `interest_rate` for the `loan50` data set. The distribution's mean is shown as a red triangle.

¹Answers 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.

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

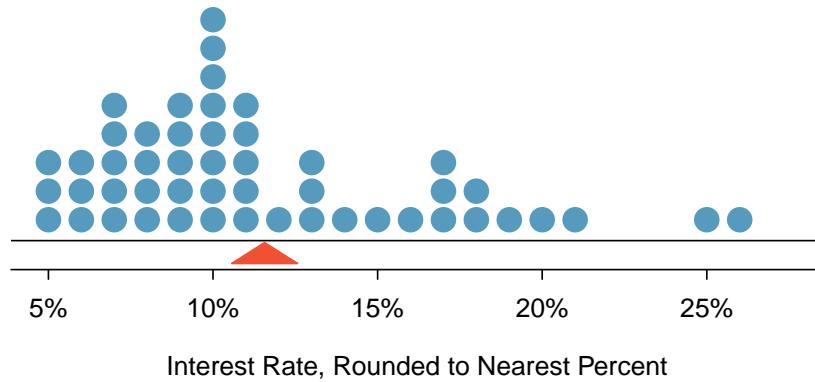


Figure 2.4: A stacked dot plot of `interest_rate` for the `loan50` data set. The rates have been rounded to the nearest percent in this plot, and the distribution's mean is shown as a red triangle.

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

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

The sample mean is often labeled \bar{x} . The letter x is being used as a generic placeholder for the variable of interest, `interest_rate`, and the bar over the x communicates we're looking at the average interest rate, which for these 50 loans was 11.57%. It is useful to think of the mean as the balancing point of the distribution, and it's shown as a triangle in Figures 2.3 and 2.4.

MEAN

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

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

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

GUIDED PRACTICE 2.4

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

GUIDED PRACTICE 2.5

What was n in this sample of loans?⁴

The `loan50` data set represents a sample from a larger population of loans made through Lending Club. We could compute a mean for this population in the same way as the sample mean. However, the population mean has a special label: μ . The symbol μ is the Greek letter *mu* and represents the average of all observations in the population. Sometimes a subscript, such as $_x$, is used to represent which variable the population mean refers to, e.g. μ_x . Often times it is too expensive to measure the population mean precisely, so we often estimate μ using the sample mean, \bar{x} .

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

⁴The sample size was $n = 50$.

EXAMPLE 2.6

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

(E)

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

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

EXAMPLE 2.7

The mean is useful because it allows us to rescale or standardize a metric into something more easily interpretable and comparable. Provide 2 examples where the mean is useful for making comparisons.

1. We would like to understand if a new drug is more effective at treating asthma attacks than the standard drug. A trial of 1500 adults is set up, where 500 receive the new drug, and 1000 receive a standard drug in the control group:

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

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

$$\text{New drug: } 200/500 = 0.4$$

$$\text{Standard drug: } 300/1000 = 0.3$$

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

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

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

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

EXAMPLE 2.8

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

(E)

The `county` data set is special in that each county actually represents many individual people. If we were to simply average across the `income` variable, we would be treating counties with 5,000 and 5,000,000 residents equally in the calculations. Instead, we should compute the total income for each county, add up all the counties' totals, and then divide by the number of people in all the counties. If we completed these steps with the `county` data, we would find that the per capita income for the US is \$30,861. Had we computed the *simple* mean of per capita income across counties, the result would have been just \$26,093!

This example used what is called a **weighted mean**. For more information on this topic, check out the following online supplement regarding weighted means openintro.org/d?file=stat_wtd_mean.

2.1.3 Histograms and shape

Dot plots show the exact value for each observation. This is useful for small data sets, but they can become hard to read with larger samples. Rather than showing the value of each observation, we prefer to think of the value as belonging to a *bin*. For example, in the `loan50` data set, we created a table of counts for the number of loans with interest rates between 5.0% and 7.5%, then the number of loans with rates between 7.5% and 10.0%, and so on. Observations that fall on the boundary of a bin (e.g. 10.00%) are allocated to the lower bin. This tabulation is shown in Figure 2.5. These binned counts are plotted as bars in Figure 2.6 into what is called a **histogram**, which resembles a more heavily binned version of the stacked dot plot shown in Figure 2.4.

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

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

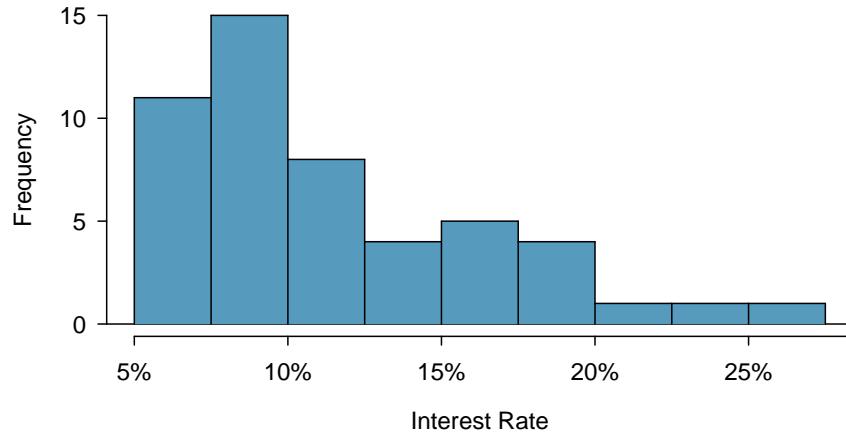


Figure 2.6: A histogram of `interest_rate`. This distribution is strongly skewed to the right.

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

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

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

LONG TAILS TO IDENTIFY SKEW

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

⁵Other ways to describe data that are right skewed: **skewed to the right**, **skewed to the high end**, or **skewed to the positive end**.

GUIDED PRACTICE 2.9

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

GUIDED PRACTICE 2.10

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

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

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

Figure 2.7 shows histograms that have one, two, or three prominent peaks. Such distributions are called **unimodal**, **bimodal**, and **multimodal**, respectively. Any distribution with more than 2 prominent peaks is called multimodal. Notice that there was one prominent peak in the unimodal distribution with a second less prominent peak that was not counted since it only differs from its neighboring bins by a few observations.

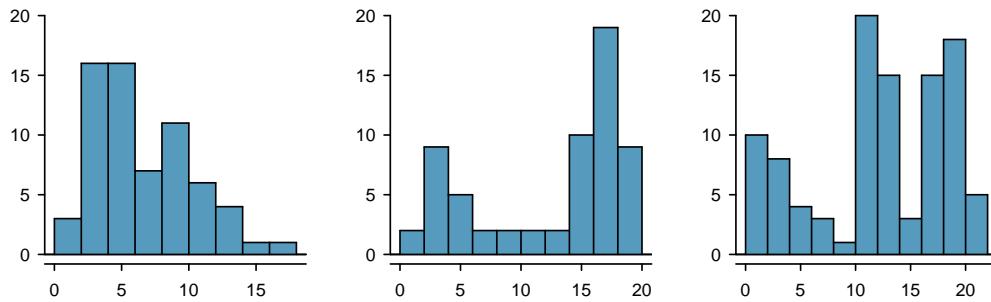


Figure 2.7: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal. Note that we've said the left plot is unimodal intentionally. This is because we are counting *prominent* peaks, not just any peak.

EXAMPLE 2.11

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

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

GUIDED PRACTICE 2.12

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

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

⁶The skew is visible in all three plots, though the flat dot plot is the least useful. The stacked dot plot and histogram are helpful visualizations for identifying skew.

⁷The interest rates for individual loans.

⁸There 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.

2.1.4 Variance and standard deviation

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

We call the distance of an observation from its mean its **deviation**. Below are the deviations for the 1st, 2nd, 3rd, and 50th observations in the `interest_rate` variable:

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

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

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

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

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

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

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

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

VARIANCE AND STANDARD DEVIATION

The variance is the average squared distance from the mean. The standard deviation is the square root of the variance. The standard deviation is useful when considering how far the data are distributed from the mean.

The standard deviation represents the typical deviation of observations from the mean. Usually about 70% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, as seen in Figures 2.8 and 2.9, these percentages are not strict rules.

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

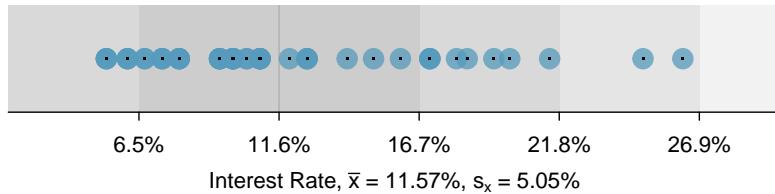


Figure 2.8: For the `interest_rate` variable, 34 of the 50 loans (68%) had interest rates within 1 standard deviation of the mean, and 48 of the 50 loans (96%) had rates within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% within 2 standard deviations,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

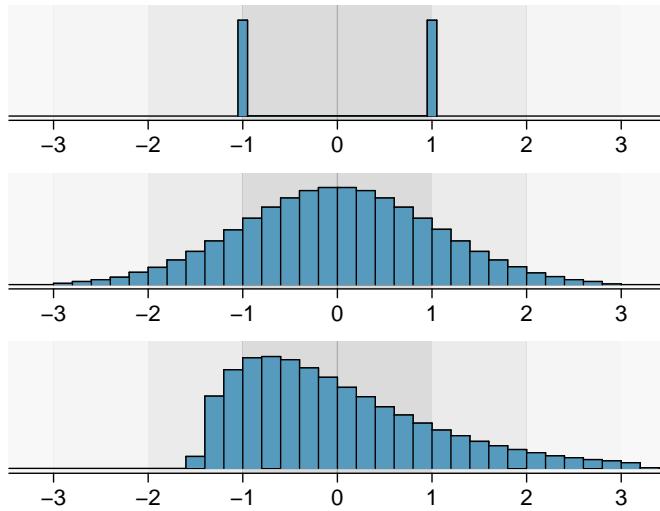


Figure 2.9: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

GUIDED PRACTICE 2.13

(G) On page 45, the concept of shape of a distribution was introduced. A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using Figure 2.9 as an example, explain why such a description is important.⁹

EXAMPLE 2.14

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

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

In practice, the variance and standard deviation are sometimes used as a means to an end, where the “end” is being able to accurately estimate the uncertainty associated with a sample statistic. For example, in Chapter 5 the standard deviation is used in calculations that help us understand how much a sample mean varies from one sample to the next.

⁹Figure 2.9 shows three distributions that look quite different, but all have the same mean, variance, and standard deviation. Using modality, we can distinguish between the first plot (bimodal) and the last two (unimodal). Using skewness, we can distinguish between the last plot (right skewed) and the first two. While a picture, like a histogram, tells a more complete story, we can use modality and shape (symmetry/skew) to characterize basic information about a distribution.

2.1.5 Box plots, quartiles, and the median

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

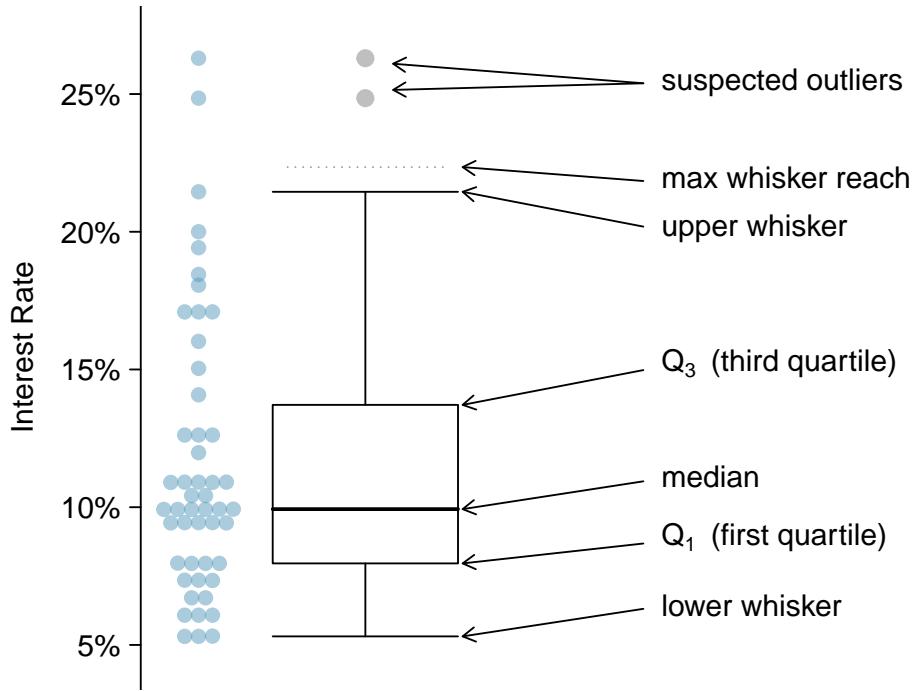


Figure 2.10: A vertical dot plot, where points have been horizontally stacked, next to a labeled box plot for the interest rates of the 50 loans.

The first step in building a box plot is drawing a dark line denoting the **median**, which splits the data in half. Figure 2.10 shows 50% of the data falling below the median and other 50% falling above the median. There are 50 loans in the data set (an even number) so the data are perfectly split into two groups of 25. We take the median in this case to be the average of the two observations closest to the 50th percentile, which happen to be the same value in this data set: $(9.93\% + 9.93\%)/2 = 9.93\%$. When there are an odd number of observations, there will be exactly one observation that splits the data into two halves, and in such a case that observation is the median (no average needed).

MEDIAN: THE NUMBER IN THE MIDDLE

If the data are ordered from smallest to largest, the **median** is the observation right in the middle. If there are an even number of observations, there will be two values in the middle, and the median is taken as their average.

The second step in building a box plot is drawing a rectangle to represent the middle 50% of the data. The total length of the box, shown vertically in Figure 2.10, is called the **interquartile range** (IQR, for short). It, like the standard deviation, is a measure of variability in data. The more variable the data, the larger the standard deviation and IQR tend to be. The two boundaries of the box are called the **first quartile** (the 25th percentile, i.e. 25% of the data fall below this value) and the **third quartile** (the 75th percentile), and these are often labeled Q_1 and Q_3 , respectively.

INTERQUARTILE RANGE (IQR)

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

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 are the 25th and 75th percentiles.

GUIDED PRACTICE 2.15

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

Extending out from the box, the **whiskers** attempt to capture the data outside of the box. However, their reach is never allowed to be more than $1.5 \times IQR$. They capture everything within this reach. In Figure 2.10, the upper whisker does not extend to the last two points, which is beyond $Q_3 + 1.5 \times IQR$, and so it extends only to the last point below this limit. The lower whisker stops at the lowest value, 5.31%, since there is no additional data to reach; the lower whisker's limit is not shown in the figure because the plot does not extend down to $Q_1 - 1.5 \times IQR$. In a sense, the box is like the body of the box plot and the whiskers are like its arms trying to reach the rest of the data.

Any observation lying beyond the whiskers is labeled with a dot. The purpose of labeling these points – instead of extending the whiskers to the minimum and maximum observed values – is to help identify any observations that appear to be unusually distant from the rest of the data. Unusually distant observations are called **outliers**. In this case, it would be reasonable to classify the interest rates of 24.85% and 26.30% as outliers since they are numerically distant from most of the data.

OUTLIERS ARE EXTREME

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

Examining data for outliers serves many useful purposes, including

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

GUIDED PRACTICE 2.16

(G) Using Figure 2.10, estimate the following values for `interest_rate` in the `loan50` data set:

(a) Q_1 , (b) Q_3 , and (c) IQR.¹¹

¹⁰Since Q_1 and Q_3 capture the middle 50% of the data and the median splits the data in the middle, 25% of the data fall between Q_1 and the median, and another 25% falls between the median and Q_3 .

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

2.1.6 Robust statistics

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

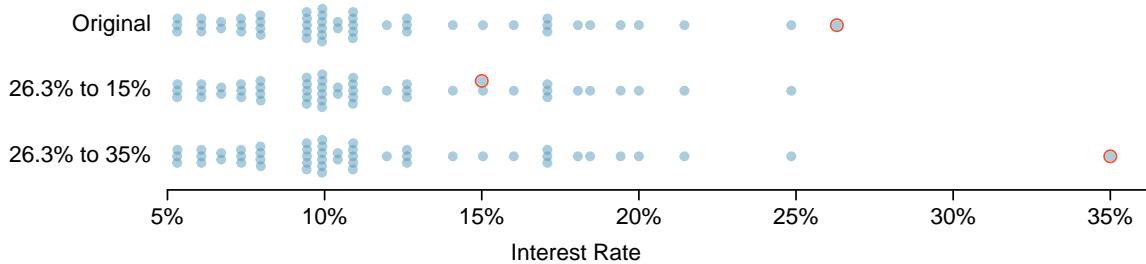


Figure 2.11: Dot plots of the original interest rate data and two modified data sets.

scenario	robust		not robust	
	median	IQR	\bar{x}	s
original <code>interest_rate</code> data	9.93%	5.76%	11.57%	5.05%
move 26.3% → 15%	9.93%	5.76%	11.34%	4.61%
move 26.3% → 35%	9.93%	5.76%	11.74%	5.68%

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

GUIDED PRACTICE 2.17

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

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

EXAMPLE 2.18

The median and IQR did not change under the three scenarios in Figure 2.12. Why might this be the case?

The median and IQR are only sensitive to numbers near Q_1 , the median, and Q_3 . Since values in these regions are stable in the three data sets, the median and IQR estimates are also stable.

GUIDED PRACTICE 2.19

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

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

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

2.1.7 Transforming data (special topic)

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

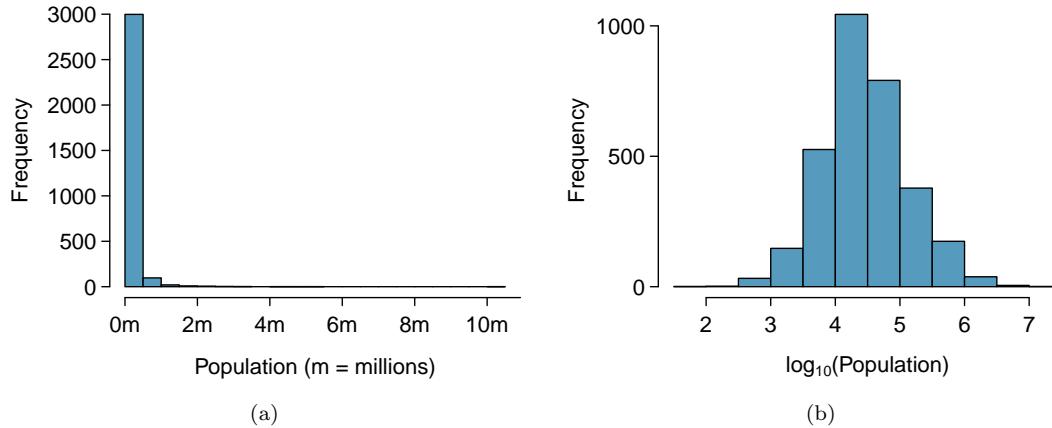


Figure 2.13: (a) A histogram of the populations of all US counties. (b) A histogram of \log_{10} -transformed county populations. For this plot, the x-value corresponds to the power of 10, e.g. “4” on the x-axis corresponds to $10^4 = 10,000$.

EXAMPLE 2.20

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

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

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

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

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

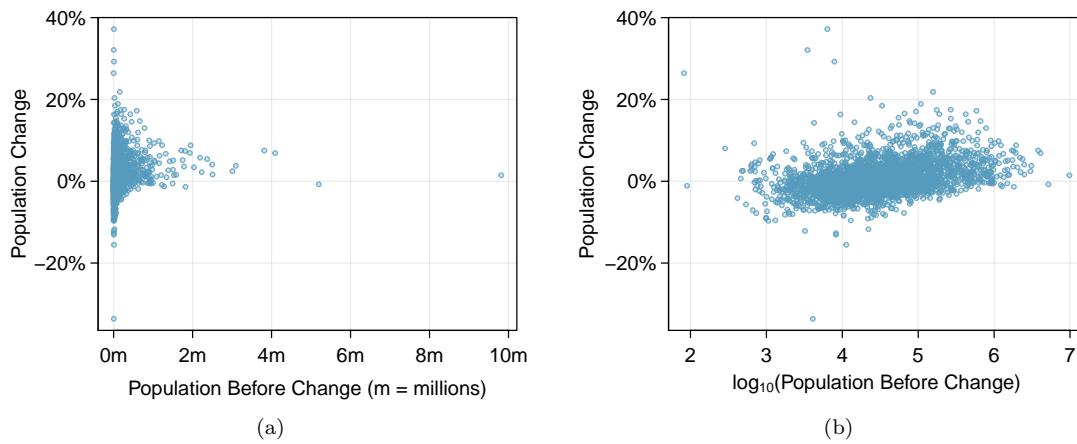


Figure 2.14: (a) Scatterplot of population change against the population before the change. (b) A scatterplot of the same data but where the population size has been log-transformed.

2.1.8 Mapping data (special topic)

The county data set offers many numerical variables that we could plot using dot plots, scatter-plots, or box plots, but these miss the true nature of the data. Rather, when we encounter geographic data, we should create an **intensity map**, where colors are used to show higher and lower values of a variable. Figures 2.15 and 2.16 shows intensity maps for poverty rate in percent (`poverty`), unemployment rate (`unemployment_rate`), homeownership rate in percent (`homeownership`), and median household income (`median_hh_income`). The color key indicates which colors correspond to which values. The intensity maps are not generally very helpful for getting precise values in any given county, but they are very helpful for seeing geographic trends and generating interesting research questions or hypotheses.

EXAMPLE 2.21

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

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

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

GUIDED PRACTICE 2.22

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

¹⁴Note: 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.

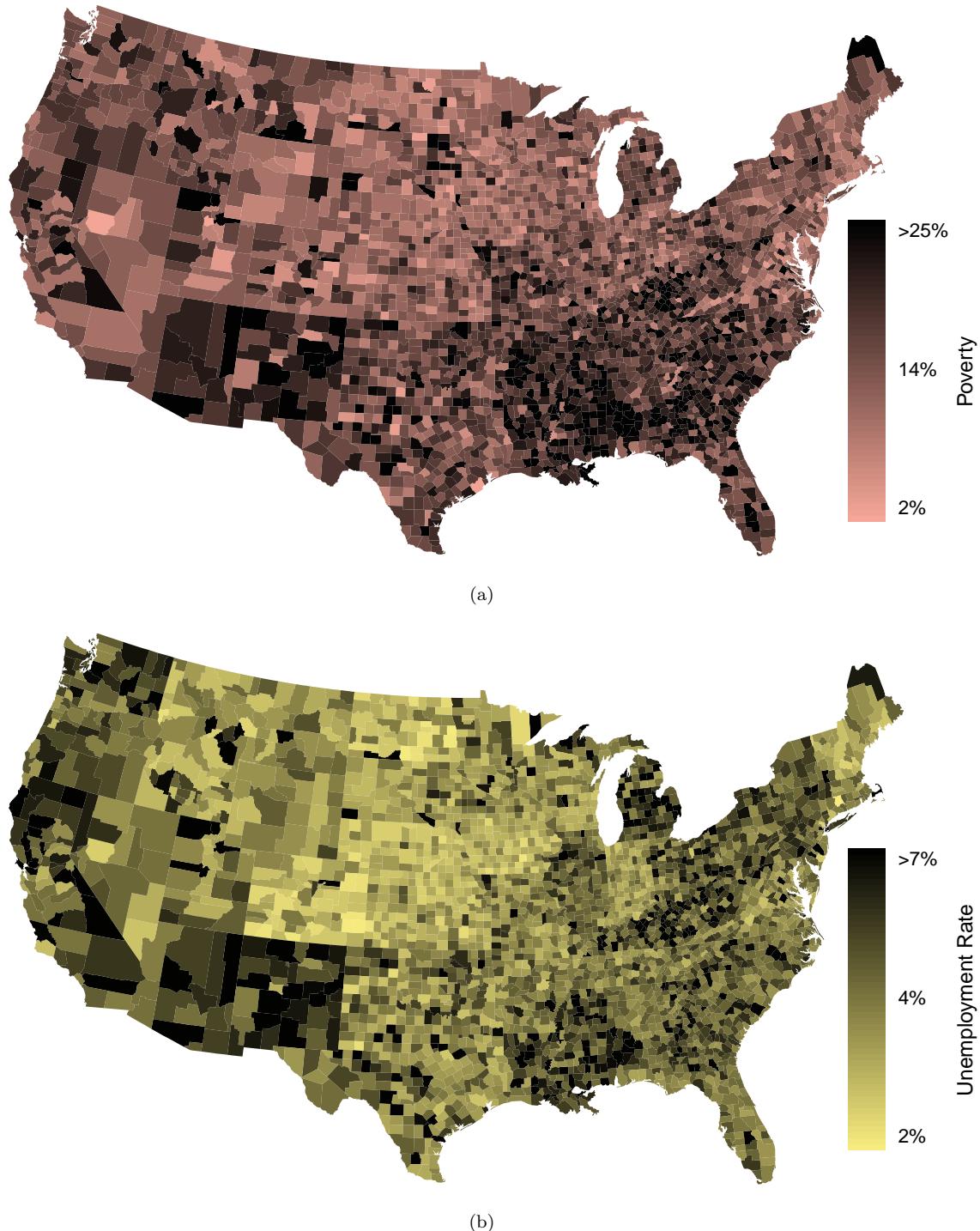


Figure 2.15: (a) Intensity map of poverty rate (percent). (b) Map of the unemployment rate (percent).

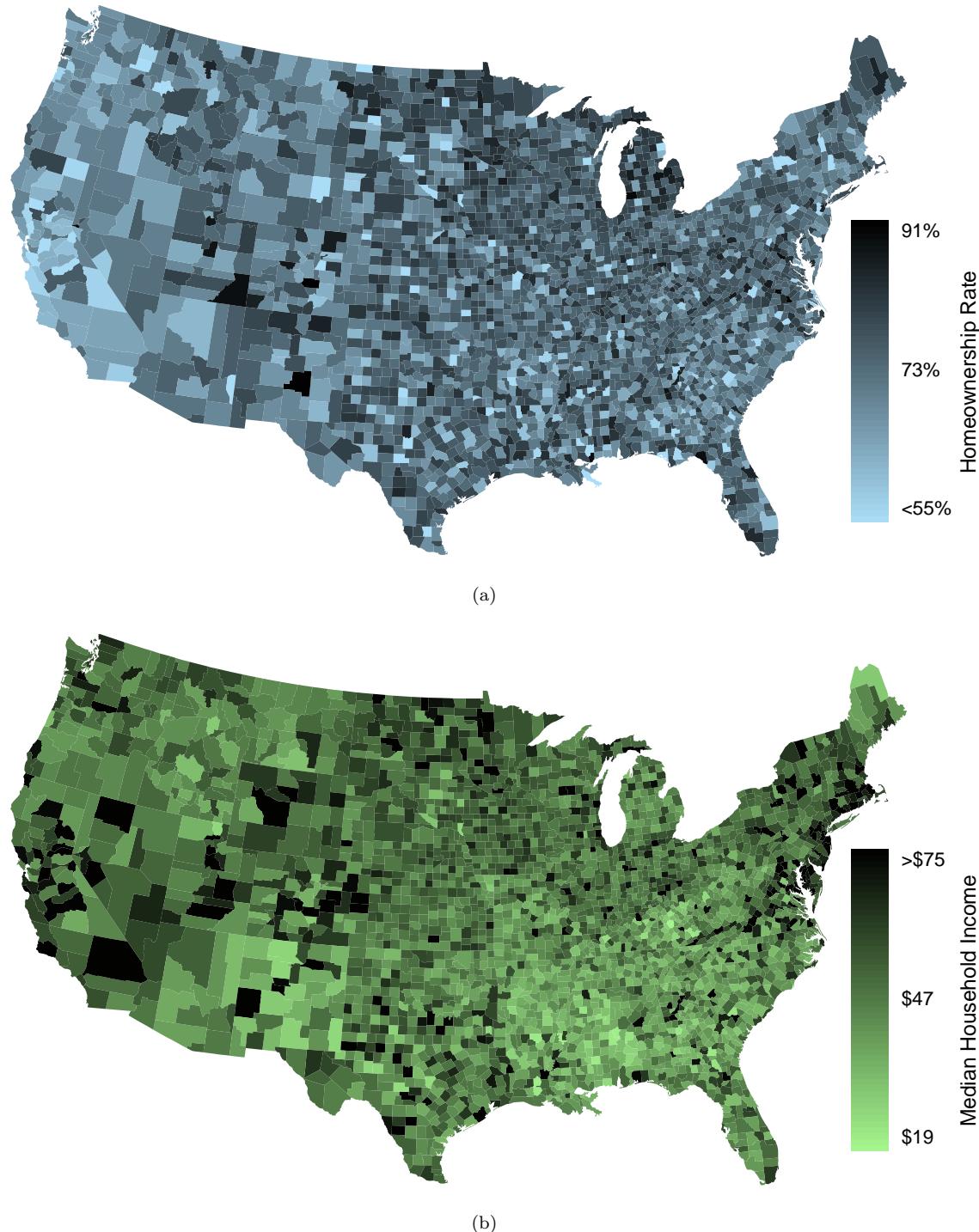
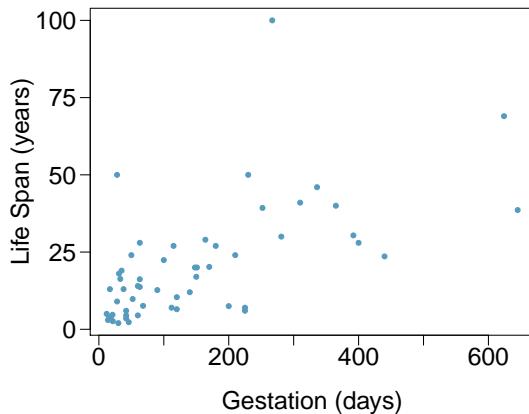


Figure 2.16: (a) Intensity map of homeownership rate (percent). (b) Intensity map of median household income (\$1000s).

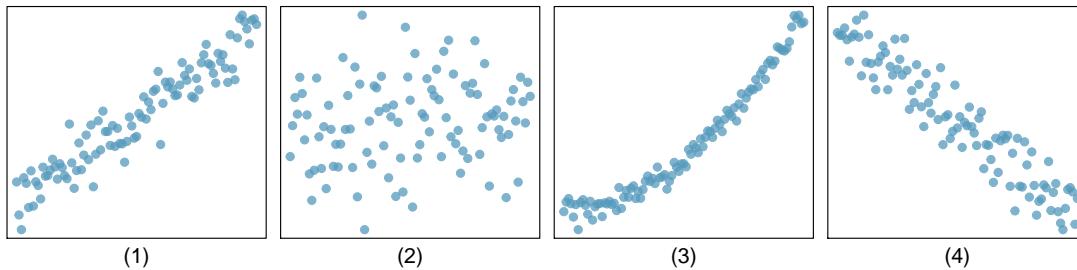
Exercises

2.1 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.¹⁵

- (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.



2.2 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.



2.3 Reproducing bacteria. Suppose that there is only sufficient space and nutrients to support one million bacterial cells in a petri dish. You place a few bacterial cells in this petri dish, allow them to reproduce freely, and record the number of bacterial cells in the dish over time. Sketch a plot representing the relationship between number of bacterial cells and time.

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

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

- (a) American households spent an average of about \$52 in 2007 on Halloween merchandise such as costumes, decorations and candy. To see if this number had changed, researchers conducted a new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was \$58 per household.
- (b) The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 a decade later.

2.6 Sleeping in college. A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night. Identify which value represents the sample mean and which value represents the claimed population mean.

¹⁵T. Allison and D.V. Cicchetti. "Sleep in mammals: ecological and constitutional correlates". In: *Arch. Hydrobiol* 75 (1975), p. 442.

2.7 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?

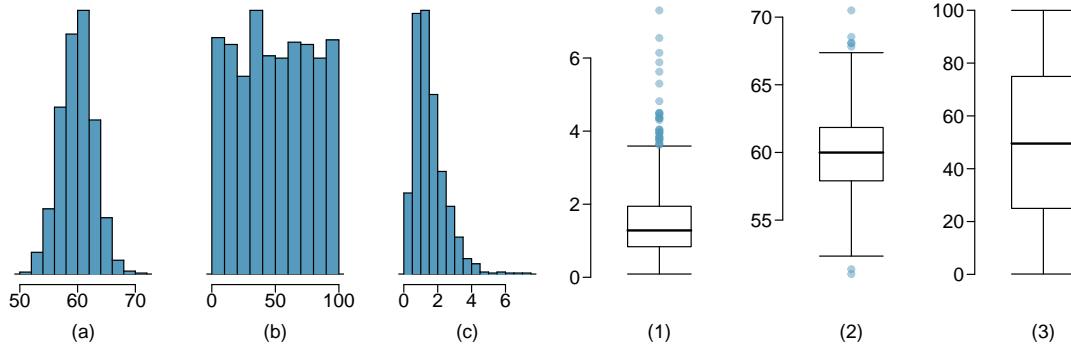
2.8 Medians and IQRs. For each part, compare distributions (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning.

- | | |
|---|--|
| (a) (1) 3, 5, 6, 7, 9
(2) 3, 5, 6, 7, 20 | (c) (1) 1, 2, 3, 4, 5
(2) 6, 7, 8, 9, 10 |
| (b) (1) 3, 5, 6, 7, 9
(2) 3, 5, 7, 8, 9 | (d) (1) 0, 10, 50, 60, 100
(2) 0, 100, 500, 600, 1000 |

2.9 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.

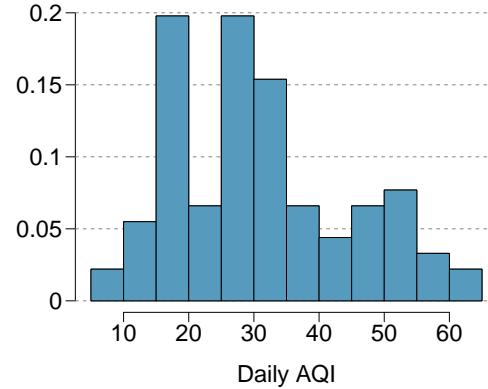
- | | |
|--|---|
| (a) (1) 3, 5, 5, 5, 8, 11, 11, 11, 13
(2) 3, 5, 5, 5, 8, 11, 11, 11, 20 | (c) (1) 0, 2, 4, 6, 8, 10
(2) 20, 22, 24, 26, 28, 30 |
| (b) (1) -20, 0, 0, 0, 15, 25, 30, 30
(2) -40, 0, 0, 0, 15, 25, 30, 30 | (d) (1) 100, 200, 300, 400, 500
(2) 0, 50, 300, 550, 600 |

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

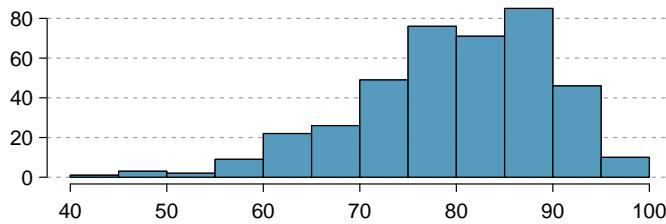


2.11 Air quality. Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air 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.¹⁶

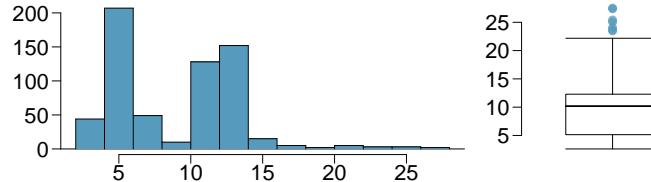
- Estimate the median AQI value of this sample.
- Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.
- Estimate Q1, Q3, and IQR for the distribution.
- Would any of the days in this sample be considered to have an unusually low or high AQI?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



2.12 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.



2.13 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?



2.14 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?¹⁷

2.15 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.

- Number of pets per household.
- Distance to work, i.e. number of miles between work and home.
- Heights of adult males.

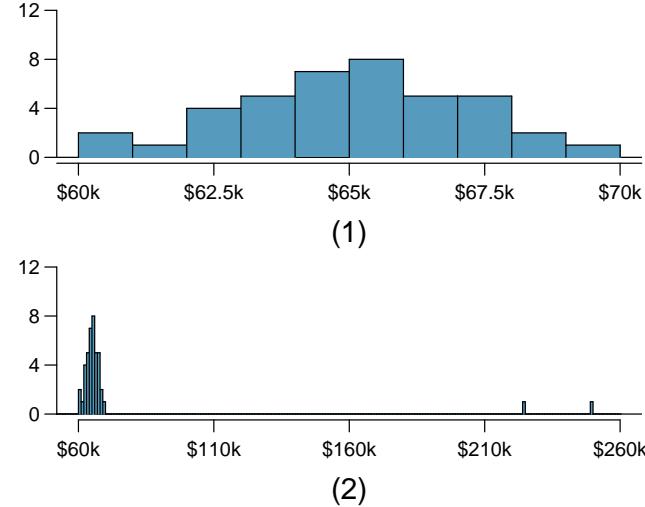
¹⁶US Environmental Protection Agency, AirData, 2011.

¹⁷Lars Backstrom. “Anatomy of Facebook”. In: *Facebook Data Team’s Notes* (2011).

2.16 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.

- 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.
- 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.
- 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.
- 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.

2.17 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.

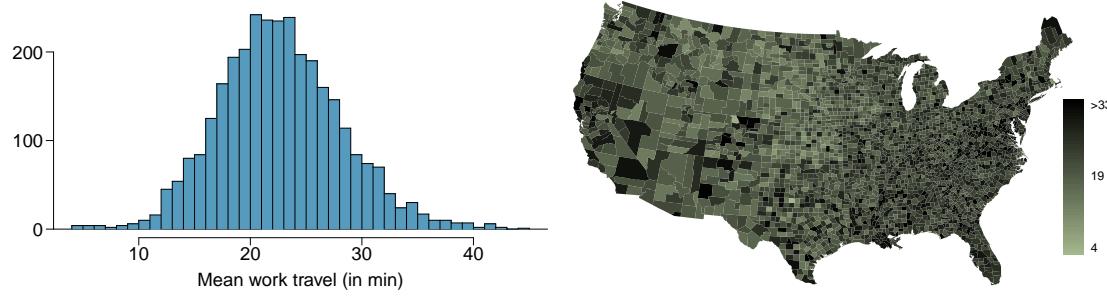


	(1)	(2)
n	40	42
Min.	60,680	60,680
1st Qu.	63,620	63,710
Median	65,240	65,350
Mean	65,090	73,300
3rd Qu.	66,160	66,540
Max.	69,890	250,000
SD	2,122	37,321

- 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?
- 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?

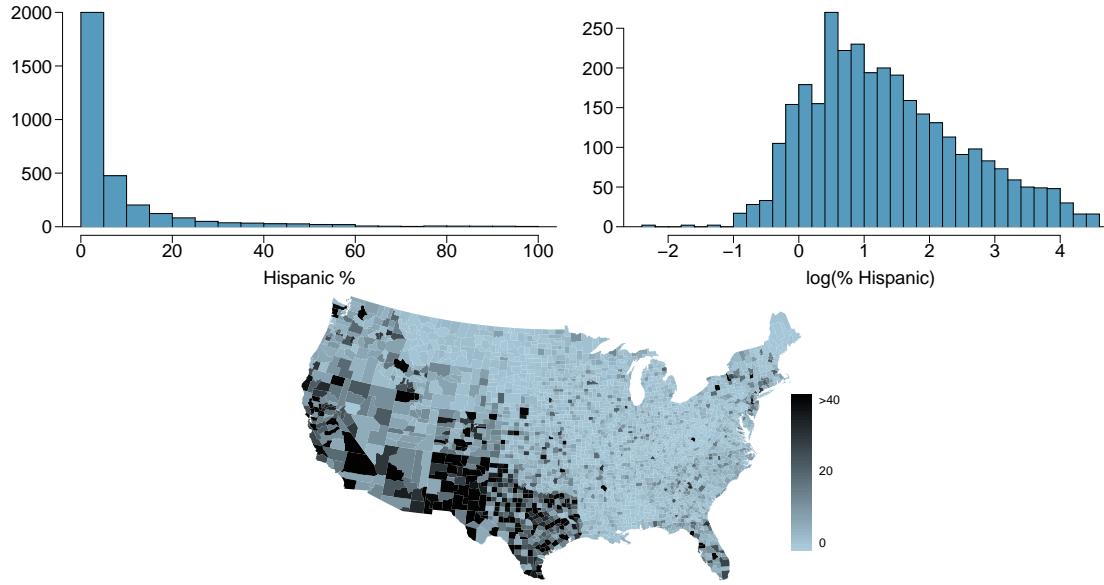
2.18 Midrange. The *midrange* of a distribution is defined as the average of the maximum and the minimum of that distribution. Is this statistic robust to outliers and extreme skew? Explain your reasoning

2.19 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 2010. Also shown below is a spatial intensity map of the same data.



- (a) Describe the numerical distribution and comment on whether or not a log transformation may be advisable for these data.
- (b) Describe the spatial distribution of commuting times using the map above.

2.20 Hispanic 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 in 3,142 counties in the US in 2010. Also shown is a histogram of logs of these values.



- (a) Describe the numerical distribution and comment on why we might want to use log-transformed values in analyzing or modeling these data.
- (b) What features of the distribution of the Hispanic 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?
- (c) Is one visualization more appropriate or helpful than the other? Explain your reasoning.

2.2 Considering categorical data

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

2.2.1 Contingency tables and bar plots

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

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

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

homeownership	Count
rent	3858
mortgage	4789
own	1353
Total	10000

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

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

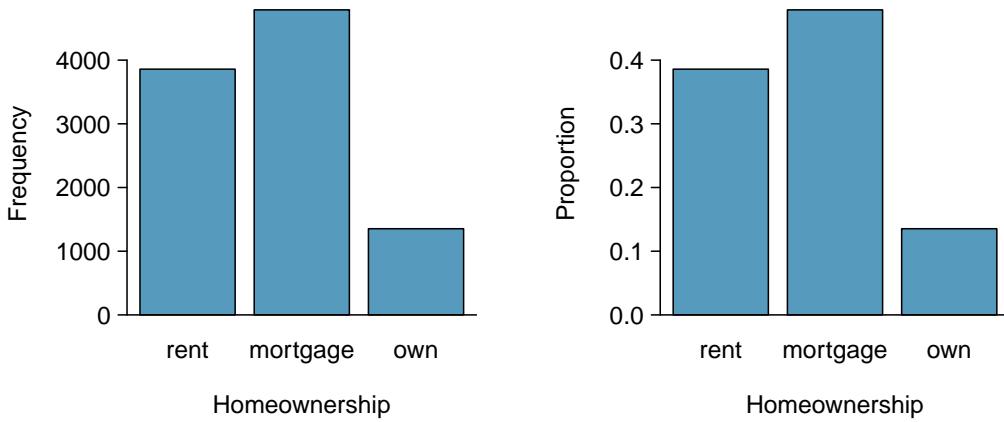


Figure 2.19: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

2.2.2 Row and column proportions

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

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

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

A contingency table of the column proportions is computed in a similar way, where each **column proportion** is computed as the count divided by the corresponding column total. Figure 2.21 shows such a table, and here the value 0.906 indicates that 90.6% of renters applied as individuals for the loan. This rate is higher compared to loans from people with mortgages (80.2%) or who own their home (86.5%). Because these rates vary between the three levels of `homeownership` (`rent`, `mortgage`, `own`), this provides evidence that the `app_type` and `homeownership` variables are associated.

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

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

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

GUIDED PRACTICE 2.23

- (G) (a) What does 0.451 represent in Figure 2.20?
 (b) What does 0.802 represent in Figure 2.21?¹⁸

GUIDED PRACTICE 2.24

- (G) (a) What does 0.122 at the intersection of `joint` and `own` represent in Figure 2.20?
 (b) What does 0.135 represent in the Figure 2.21?¹⁹

EXAMPLE 2.25

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

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

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

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

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

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

EXAMPLE 2.26

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

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

¹⁸(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.

¹⁹(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.

2.2.3 Using a bar plot with two variables

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

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

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

For the last type of bar plot we introduce, the column proportions for the `app_type` and `homeownership` contingency table have been translated into a standardized stacked bar plot in Figure 2.23(c). This type of visualization is helpful in understanding the fraction of individual or joint loan applications for borrowers in each level of `homeownership`. Additionally, since the proportions of `joint` and `individual` vary across the groups, we can conclude that the two variables are associated.

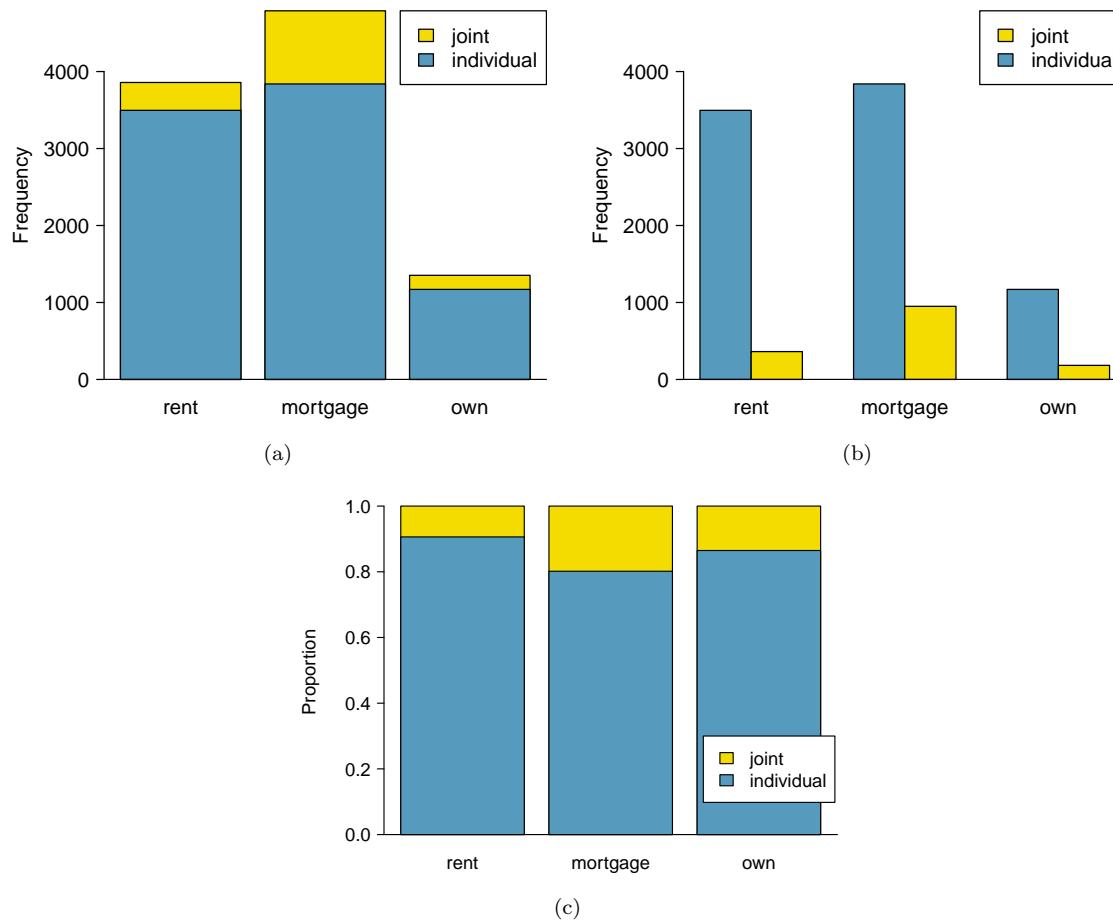


Figure 2.23: (a) Stacked bar plot for `homeownership`, where the counts have been further broken down by `app_type`. (b) Side-by-side bar plot for `homeownership` and `app_type`. (c) Standardized version of the stacked bar plot.

EXAMPLE 2.27

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

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

(E)

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

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

2.2.4 Mosaic plots

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

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

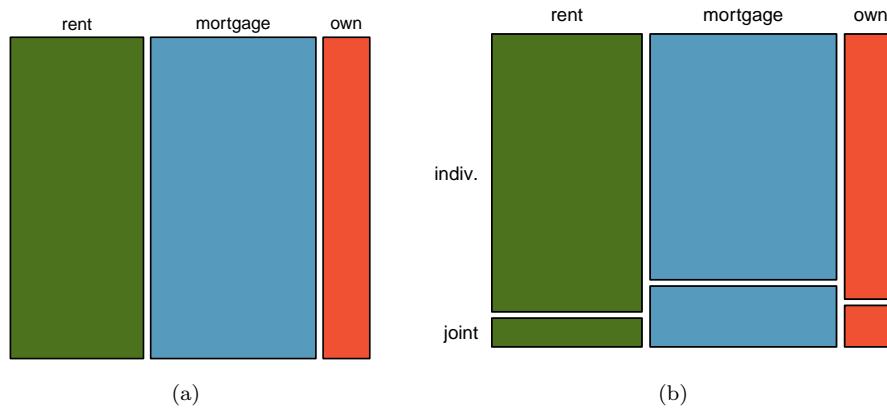


Figure 2.24: (a) The one-variable mosaic plot for `homeownership`. (b) Two-variable mosaic plot for both `homeownership` and `app_type`.

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

vertical locations than others, which was the same technique used for checking an association in the standardized stacked bar plot.

In Figure 2.24, we chose to first split by the homeowner status of the borrower. However, we could have instead first split by the application type, as in Figure 2.25. Like with the bar plots, it's common to use the explanatory variable to represent the first split in a mosaic plot, and then for the response to break up each level of the explanatory variable, if these labels are reasonable to attach to the variables under consideration.

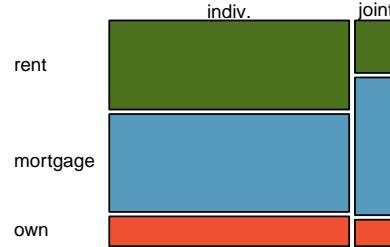


Figure 2.25: Mosaic plot where loans are grouped by the `homeownership` variable after they've been divided into the `individual` and `joint` application types.

2.2.5 The only pie chart you will see in this book

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

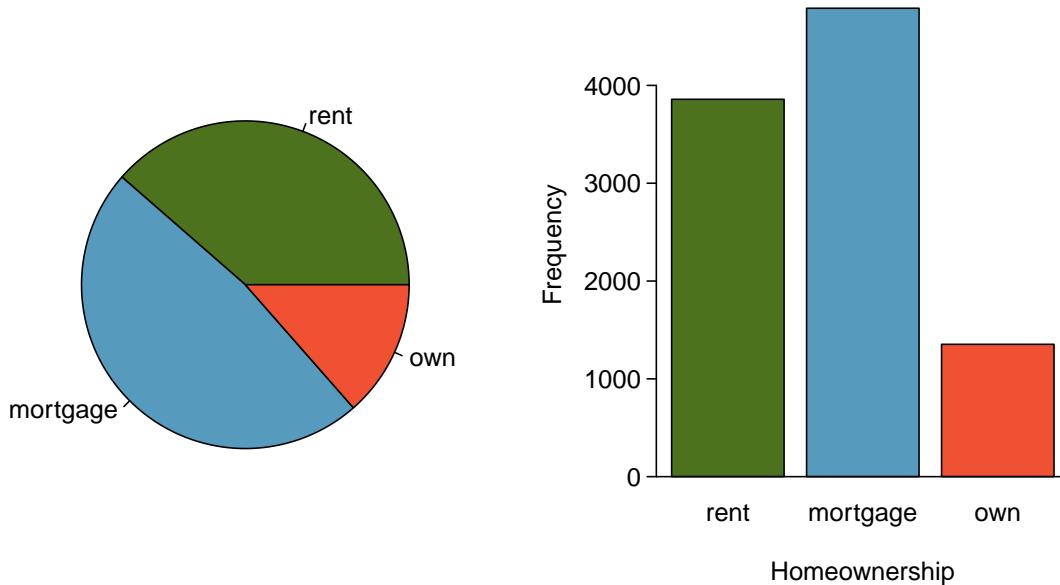


Figure 2.26: A pie chart and bar plot of `homeownership`.

2.2.6 Comparing numerical data across groups

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

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

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

Median Income for 150 Counties, in \$1000s					
Population Gain			No Population Gain		
38.2	43.6	42.2	61.5	51.1	45.7
44.6	51.8	40.7	48.1	56.4	41.9
40.6	63.3	52.1	60.3	49.8	51.7
51.1	34.1	45.5	52.8	49.1	51
80.8	46.3	82.2	43.6	39.7	49.4
75.2	40.6	46.3	62.4	44.1	51.3
51.9	34.7	54	42.9	52.2	45.1
61	51.4	56.5	62	46	46.4
53.8	57.6	69.2	48.4	40.5	48.6
53.1	54.6	55	46.4	39.9	56.7
63	49.1	57.2	44.1	50	38.9
46.6	46.5	38.9	50.9	56	34.6
74.2	63	49.6	53.7	77.5	60
63.2	47.6	55.9	39.1	57.8	42.6
50.4	49	45.6	39	38.8	37.1
57.2	44.7	71.7	35.3	100.2	
42.6	55.5	38.6	52.7	63	
					43.4
					56.5

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

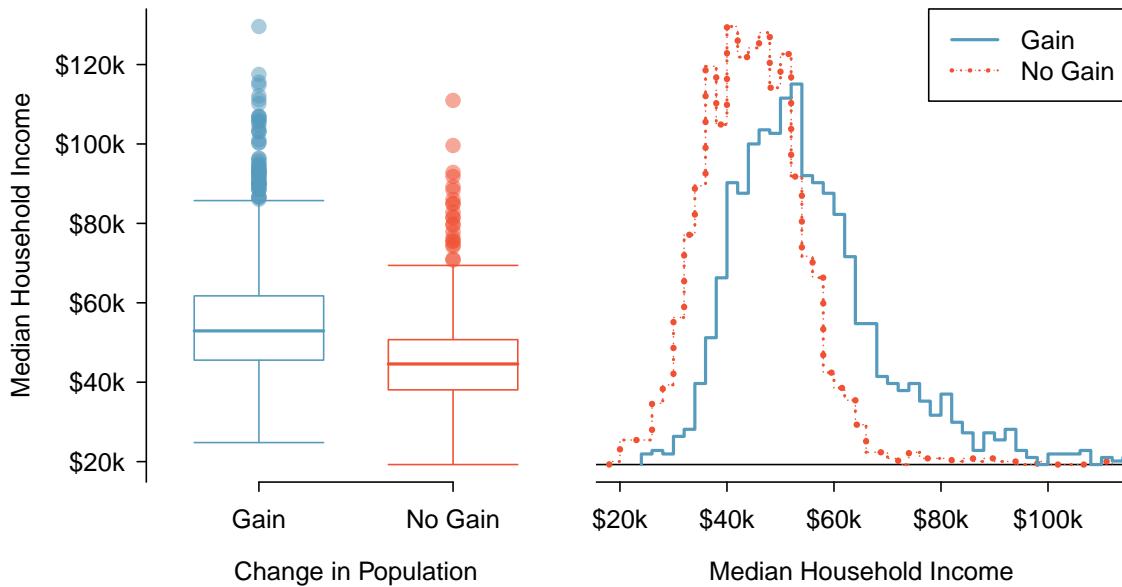


Figure 2.28: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_hh_income`, where the counties are split by whether there was a population gain or loss.

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

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

GUIDED PRACTICE 2.28

Use the plots in Figure 2.28 to compare the incomes for counties across the two groups. What do you notice about the approximate center of each group? What do you notice about the variability between groups? Is the shape relatively consistent between groups? How many *prominent* modes are there for each group?²⁰

GUIDED PRACTICE 2.29

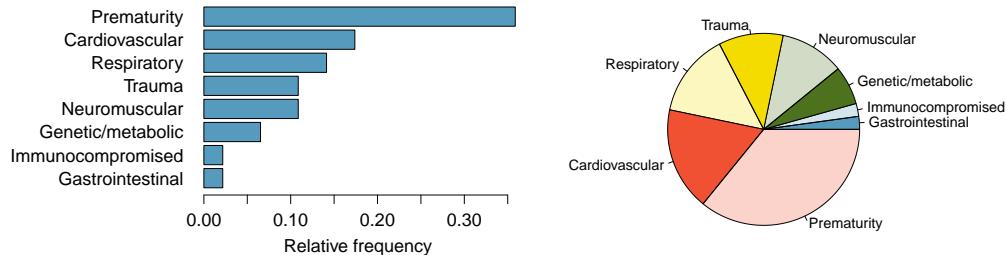
What components of each plot in Figure 2.28 do you find most useful?²¹

²⁰Answers 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.

²¹Answers will vary. The side-by-side box plots are especially useful for comparing centers and spreads, while the hollow histograms are more useful for seeing distribution shape, skew, and potential anomalies.

Exercises

2.21 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?

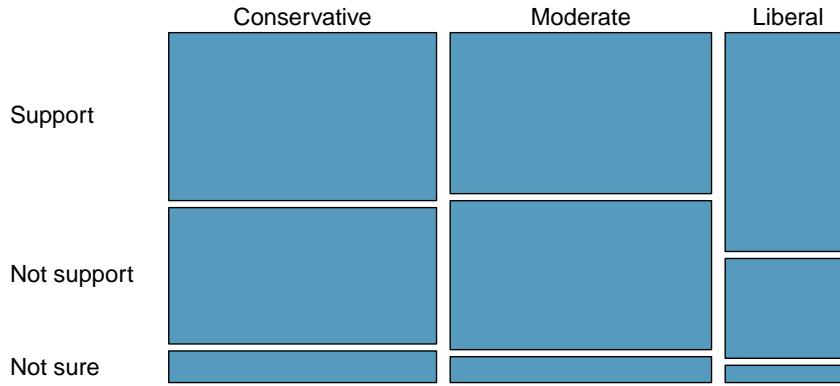
2.22 Views on immigration. 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply 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.²²

		Political ideology			Total
		Conservative	Moderate	Liberal	
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

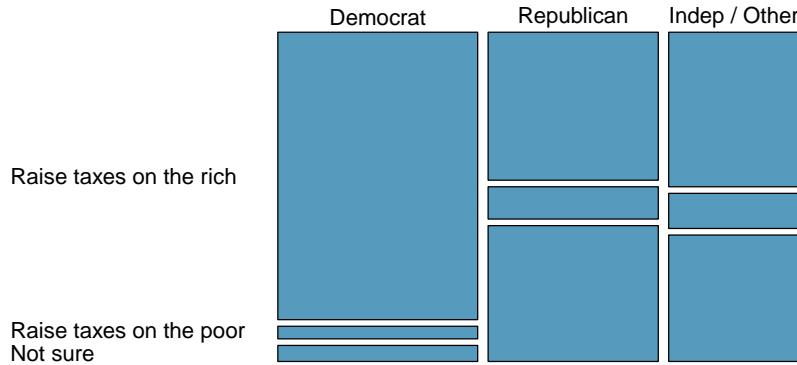
- (a) What percent of these Tampa, FL voters identify themselves as conservatives?
- (b) What percent of these Tampa, FL voters are in favor of the citizenship option?
- (c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?
- (d) What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

²²SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

2.23 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.²³



2.24 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.²⁴



²³SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

²⁴Public Policy Polling, Americans on College Degrees, Classic Literature, the Seasons, and More, data collected Feb 20-22, 2015.

2.3 Case study: malaria vaccine

EXAMPLE 2.30

Suppose your professor splits the students in class into two groups: students on the left and students on the right. If \hat{p}_L and \hat{p}_R represent the proportion of students who own an Apple product on the left and right, respectively, would you be surprised if \hat{p}_L did not exactly equal \hat{p}_R ?²⁵

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

GUIDED PRACTICE 2.31

If we don't think the side of the room a person sits on in class is related to whether the person owns an Apple product, what assumption are we making about the relationship between these two variables?²⁵

2.3.1 Variability within data

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

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

Figure 2.29: Summary results for the malaria vaccine experiment.

GUIDED PRACTICE 2.32

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

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.

²⁵We would be assuming that these two variables are independent.

²⁶The study is an experiment, as patients were randomly assigned an experiment group. Since this is an experiment, the results can be used to evaluate a causal relationship between the malaria vaccine and whether patients showed signs of an infection.

EXAMPLE 2.33

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

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

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

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

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

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

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

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

2.3.2 Simulating the study

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

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

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

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

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

GUIDED PRACTICE 2.34

(G)

What is the difference in infection rates between the two simulated groups in Figure 2.30? How does this compare to the observed 64.3% difference in the actual data?²⁷

2.3.3 Checking for independence

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

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

And another:

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

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

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

EXAMPLE 2.35

(E)

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

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

²⁷ $4/6 - 7/14 = 0.167$ or about 16.7% in favor of the vaccine. This difference due to chance is much smaller than the difference observed in the actual groups.

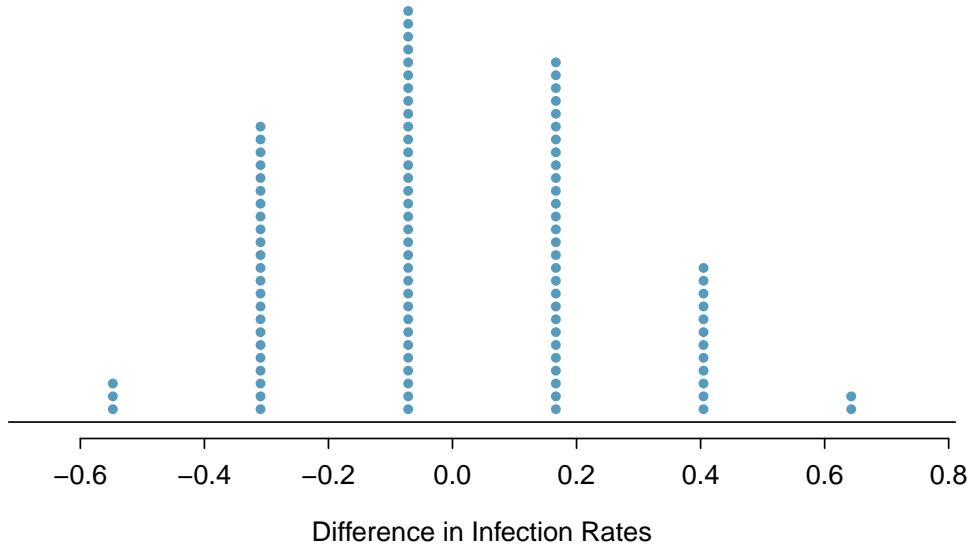


Figure 2.31: A stacked dot plot of differences from 100 simulations produced under the independence model, H_0 , where in these simulations infections are unaffected by the vaccine. Two of the 100 simulations had a difference of at least 64.3%, the difference observed in the study.

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

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

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

Based on the simulations, we have two options. (1) We conclude that the study results do not provide strong evidence against the independence model. That is, we do not have sufficiently strong evidence to conclude the vaccine had an effect in this clinical setting. (2) We conclude the evidence is sufficiently strong to reject H_0 and assert that the vaccine was useful. When we conduct formal studies, usually we reject the notion that we just happened to observe a rare event.²⁸ So in this case, we reject the independence model in favor of the alternative. That is, we are concluding the data provide strong evidence that the vaccine provides some protection against malaria in this clinical setting.

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

²⁸This reasoning does not generally extend to anecdotal observations. Each of us observes incredibly rare events every day, events we could not possibly hope to predict. However, in the non-rigorous setting of anecdotal evidence, almost anything may appear to be a rare event, so the idea of looking for rare events in day-to-day activities is treacherous. For example, we might look at the lottery: there was only a 1 in 292 million chance that the Powerball numbers for the largest jackpot in history (January 13th, 2016) would be (04, 08, 19, 27, 34) with a Powerball of (10), but nonetheless those numbers came up! However, no matter what numbers had turned up, they would have had the same incredibly rare odds. That is, *any set of numbers we could have observed would ultimately be incredibly rare*. This type of situation is typical of our daily lives: each possible event in itself seems incredibly rare, but if we consider every alternative, those outcomes are also incredibly rare. We should be cautious not to misinterpret such anecdotal evidence.

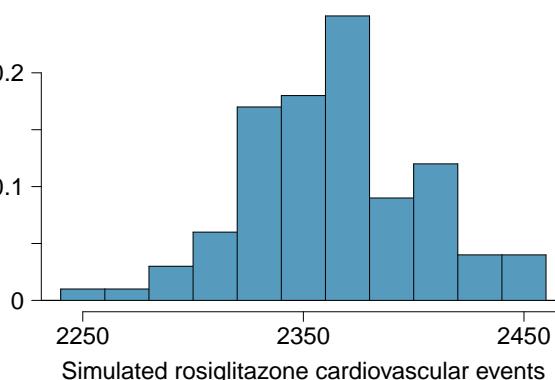
Exercises

2.25 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.²⁹

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

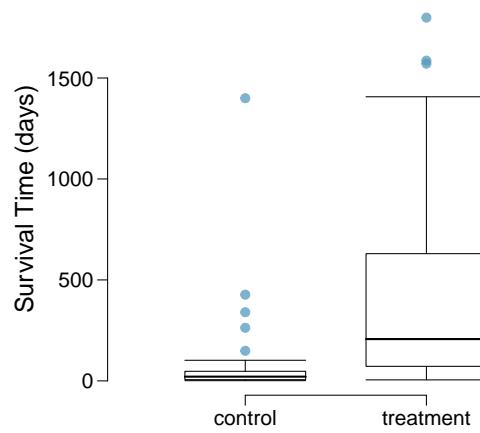
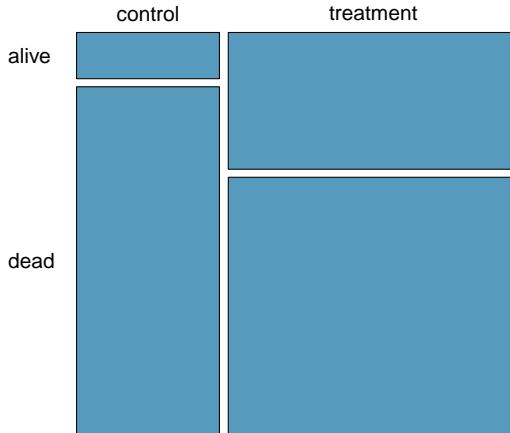
- (a) Determine if each of the following statements is true or false. If false, explain why. *Be careful:* The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.
 - i. Since more patients on pioglitazone had cardiovascular problems (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 (c), 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?



²⁹D.J. Graham et al. "Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone". In: *JAMA* 304.4 (2010), p. 411. ISSN: 0098-7484.

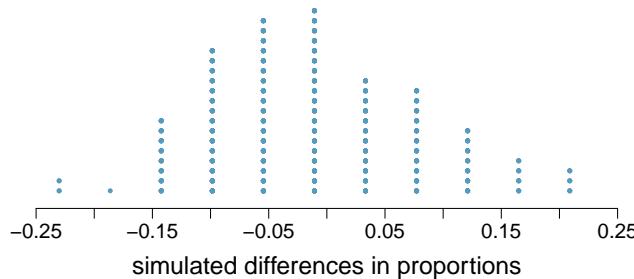
2.26 Heart transplants. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most 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.³⁰



- Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.
- What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.
- What proportion of patients in the treatment group and what proportion of patients in the control group died?
- One approach for investigating whether or not the treatment is effective is to use a randomization technique.
 - What are the claims being tested?
 - 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.

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



³⁰B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.

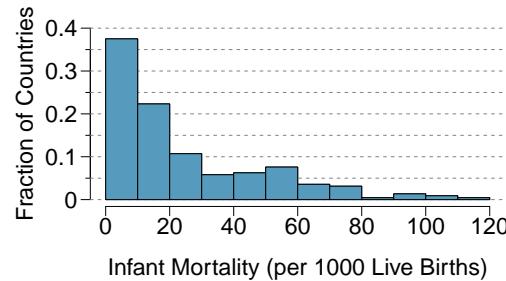
Chapter exercises

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

- Does the new student's score increase or decrease the average score?
- What is the new average?
- Does the new student's score increase or decrease the standard deviation of the scores?

2.28 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 2014.³¹

- Estimate Q1, the median, and Q3 from the histogram.
- Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.

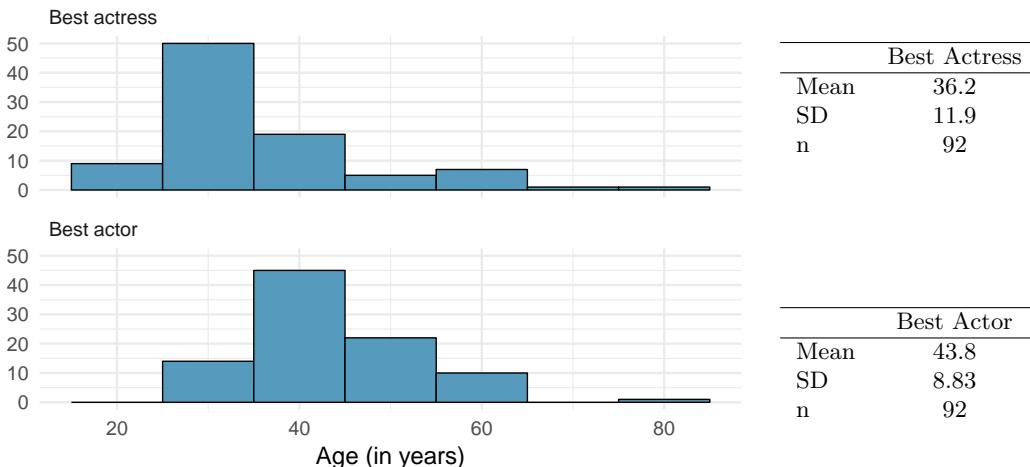


2.29 TV watchers. Students in an AP Statistics class were asked how many hours of television they watch per week (including online streaming). This sample yielded an average of 4.71 hours, with a standard deviation of 4.18 hours. Is the distribution of number of hours students watch television weekly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

2.30 A new statistic. The statistic $\frac{\bar{x}}{\text{median}}$ 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.

- $\frac{\bar{x}}{\text{median}} = 1$
- $\frac{\bar{x}}{\text{median}} < 1$
- $\frac{\bar{x}}{\text{median}} > 1$

2.31 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.³²



³¹CIA Factbook, Country Comparisons, 2014.

³²Oscar winners from 1929 – 2012, data up to 2009 from the Journal of Statistics Education data archive and more current data from wikipedia.org.

2.32 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.

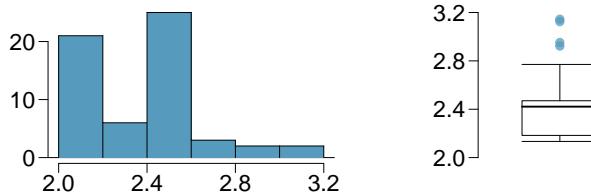
2.33 Stats scores. Below are the final exam scores of twenty introductory statistics students.

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

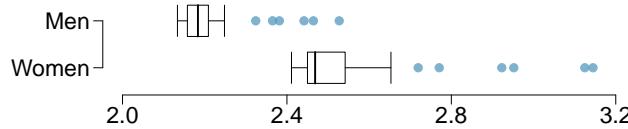
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

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

2.34 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.

