Section 2.1: Examining Numerical Data

Diez, D. M., Çetinkaya-Rundel, M., Barr, C. D. (2019). OpenIntro Statistics (4th ed.). OpenIntro. https://www.openintro.org/book/os/ CC BY-SA 3.0

STAT 1201
Introduction to Probability and Statistics

Chapter 2

Summarizing data

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.

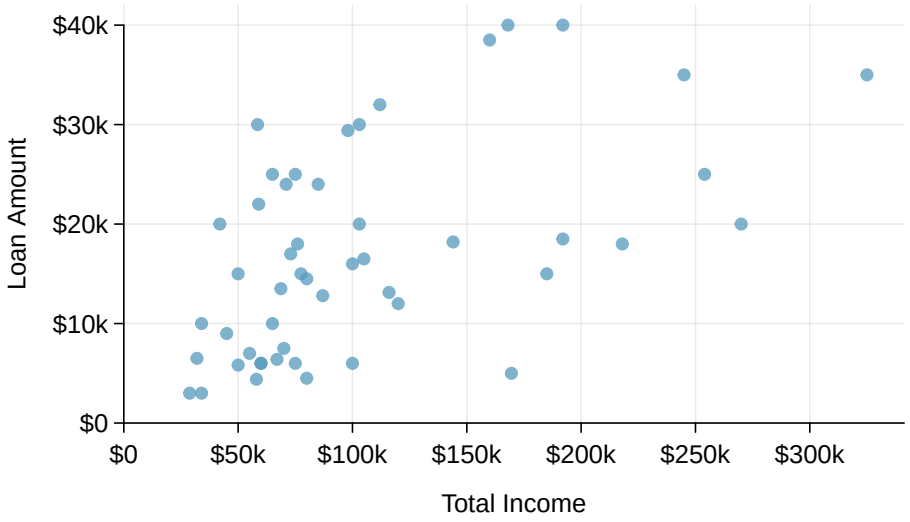


Figure 2.1: A scatterplot of `total_income` versus `loan_amount` for the `loan50` data set.

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?

(E)

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

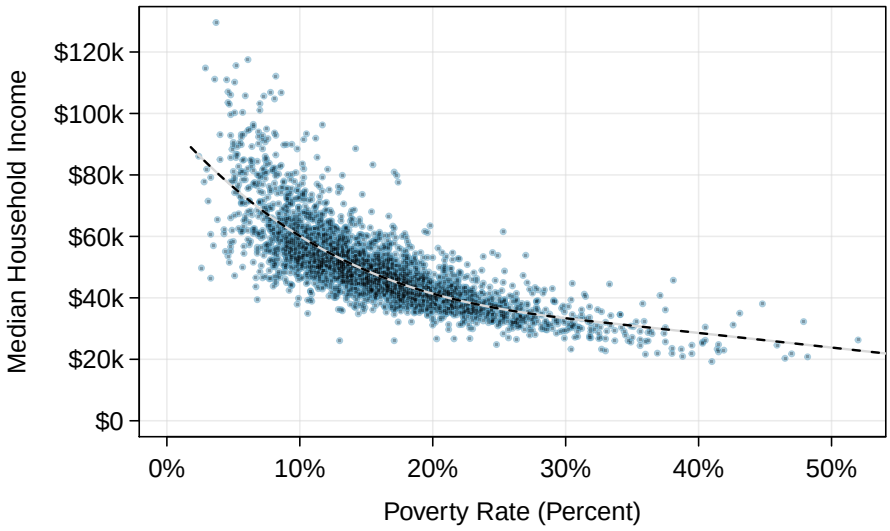


Figure 2.2: A scatterplot of the median household income against the poverty rate for the `county` data set. A statistical model has also been fit to the data and is shown as a dashed line.

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 $\frown$).[2]

2.1.2 Dot plots and the mean

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

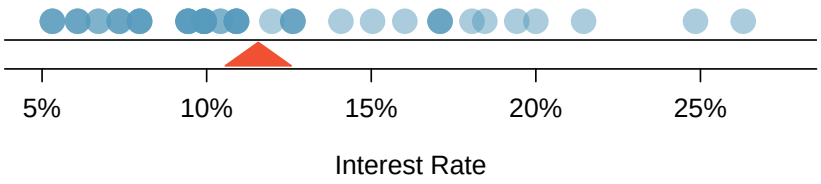


Figure 2.3: A dot plot of `interest_rate` for the `loan50` data set. The distribution's mean is shown as a red triangle.

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

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

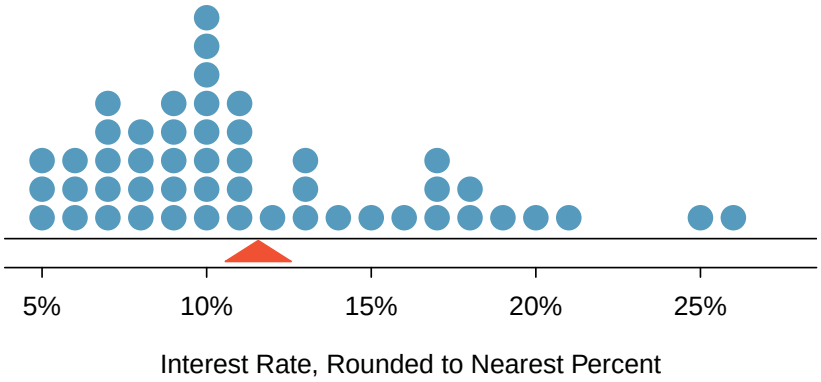


Figure 2.4: A stacked dot plot of `interest_rate` for the `loan50` data set. The rates have been rounded to the nearest percent in this plot, and the distribution's mean is shown as a red triangle.

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

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

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

MEAN

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

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

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

GUIDED PRACTICE 2.4

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

GUIDED PRACTICE 2.5

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

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

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

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

EXAMPLE 2.6

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

(E)

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

In Chapter 5 and beyond, we will develop tools to characterize the accuracy of *point estimates* like the sample mean. 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

(E)

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:

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

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.

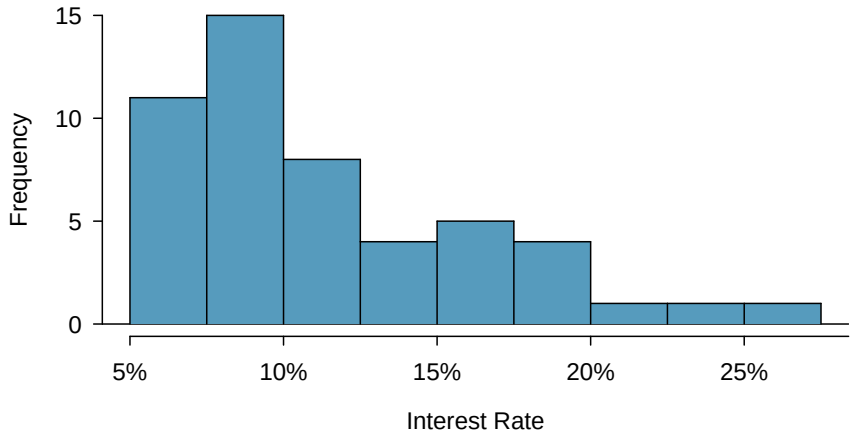


Figure 2.6: A histogram of `interest_rate`. This distribution is strongly skewed to the right.

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.

[5]Other 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.

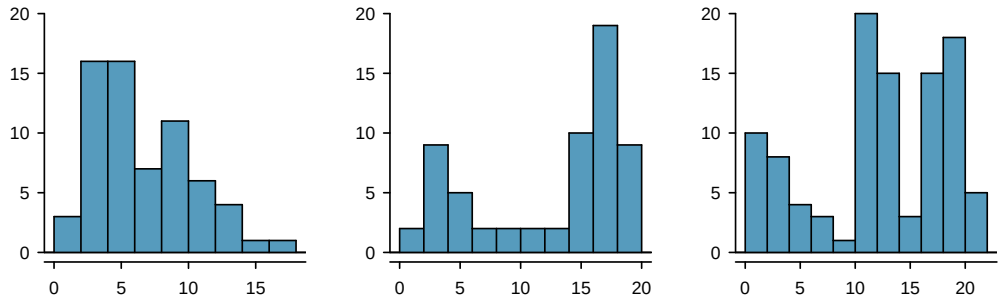


Figure 2.7: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal. Note that we've said the left plot is unimodal intentionally. This is because we are counting *prominent* peaks, not just any peak.

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 *uni*cycles). Similarly, *bi* stands for 2 (think *bi*cycles). We're hoping a *multicycle* will be invented to complete this analogy.

GUIDED PRACTICE 2.12

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

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

---

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

[7] The interest rates for individual loans.

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

2.1.4 Variance and standard deviation

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

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

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

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

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

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

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

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

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

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

VARIANCE AND STANDARD DEVIATION

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

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

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

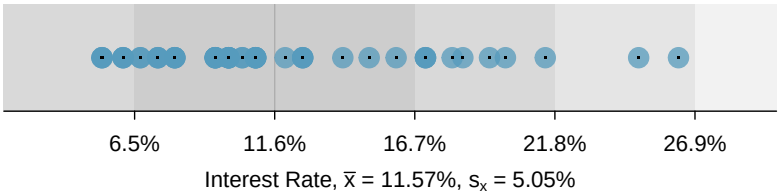


Figure 2.8: For the `interest_rate` variable, 34 of the 50 loans (68%) had interest rates within 1 standard deviation of the mean, and 48 of the 50 loans (96%) had rates within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% within 2 standard deviations,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

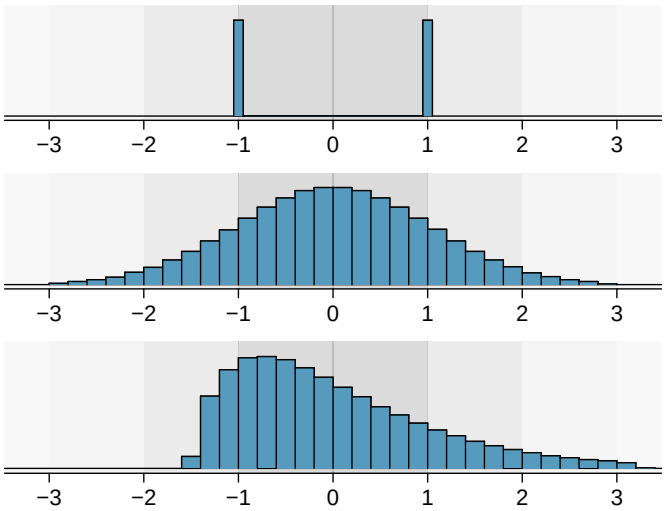


Figure 2.9: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

GUIDED PRACTICE 2.13

G On page 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

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.

E

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.

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

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 50^{th} percentile, which happen to be the same value in this data set: $(9.93\% + 9.93\%)/2 = 9.93\%$. When there are an odd number of observations, there will be exactly one observation that splits the data into two halves, and in such a case that observation is the median (no average needed).

MEDIAN: THE NUMBER IN THE MIDDLE

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

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

INTERQUARTILE RANGE (IQR)

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

$$IQR = Q_3 - Q_1$$

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

GUIDED PRACTICE 2.15

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

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

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

OUTLIERS ARE EXTREME

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

Examining data for outliers serves many useful purposes, including

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

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]

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

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

2.1.6 Robust statistics

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

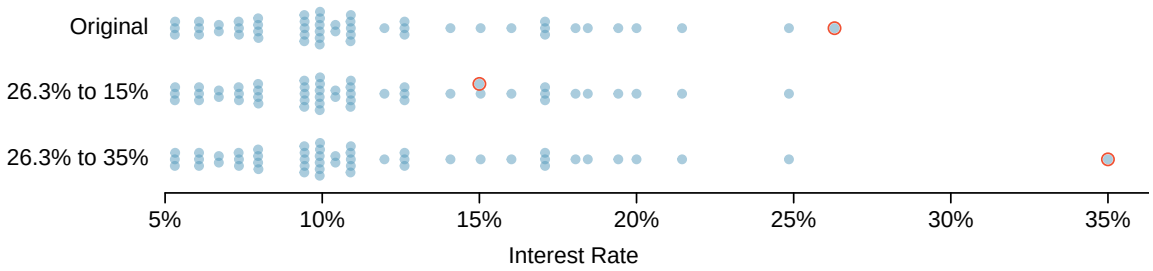


Figure 2.11: Dot plots of the original interest rate data and two modified data sets.

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

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

GUIDED PRACTICE 2.17

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

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

EXAMPLE 2.18

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

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

GUIDED PRACTICE 2.19

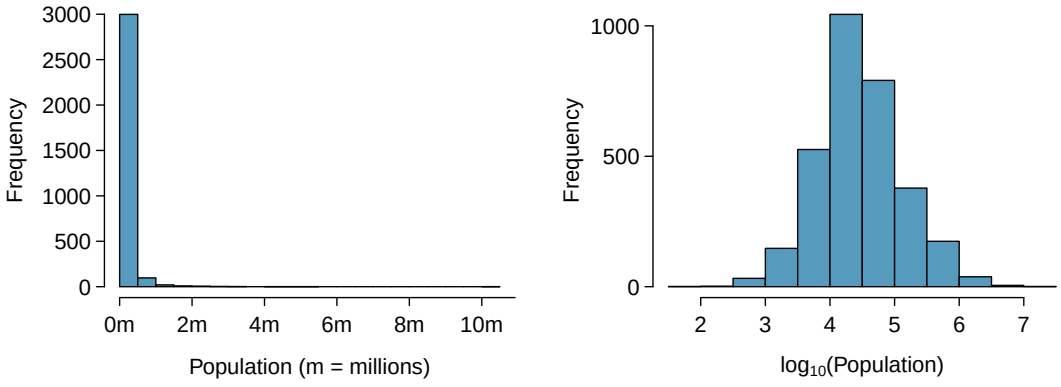
The distribution of loan amounts in the `loan50` data set is right skewed, with a few large loans lingering out into the right tail. If you were wanting to understand the typical loan size, should you be more interested in the mean or median?[13]

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

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

2.1.7 Transforming data (special topic)

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



(a) (b)

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

EXAMPLE 2.20

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

E

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

There are some standard transformations that may be useful for strongly right skewed data where much of the data is positive but clustered near zero. A **transformation** is a rescaling of the data using a function. For instance, a plot of the logarithm (base 10) of county populations results in the new histogram in Figure 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.

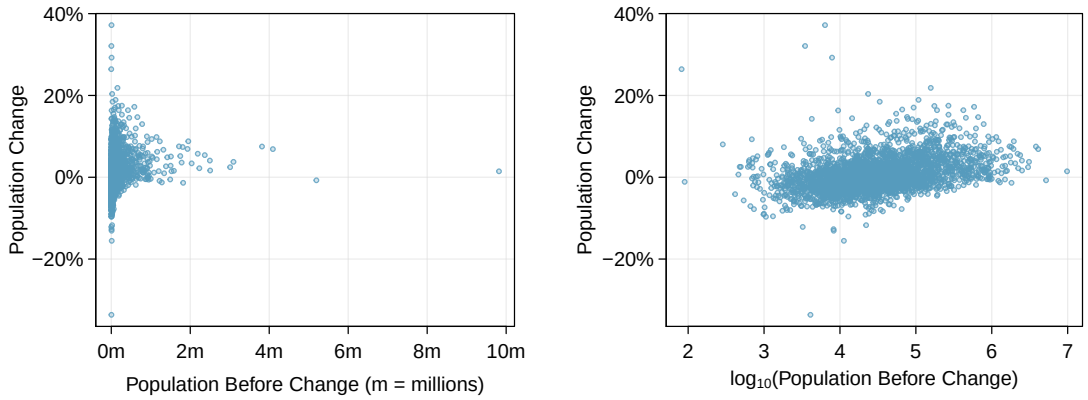


Figure 2.14: (a) Scatterplot of population change against the population before the change. (b) A scatterplot of the same data but where the population size has been log-transformed.

2.1.8 Mapping data (special topic)

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

EXAMPLE 2.21

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

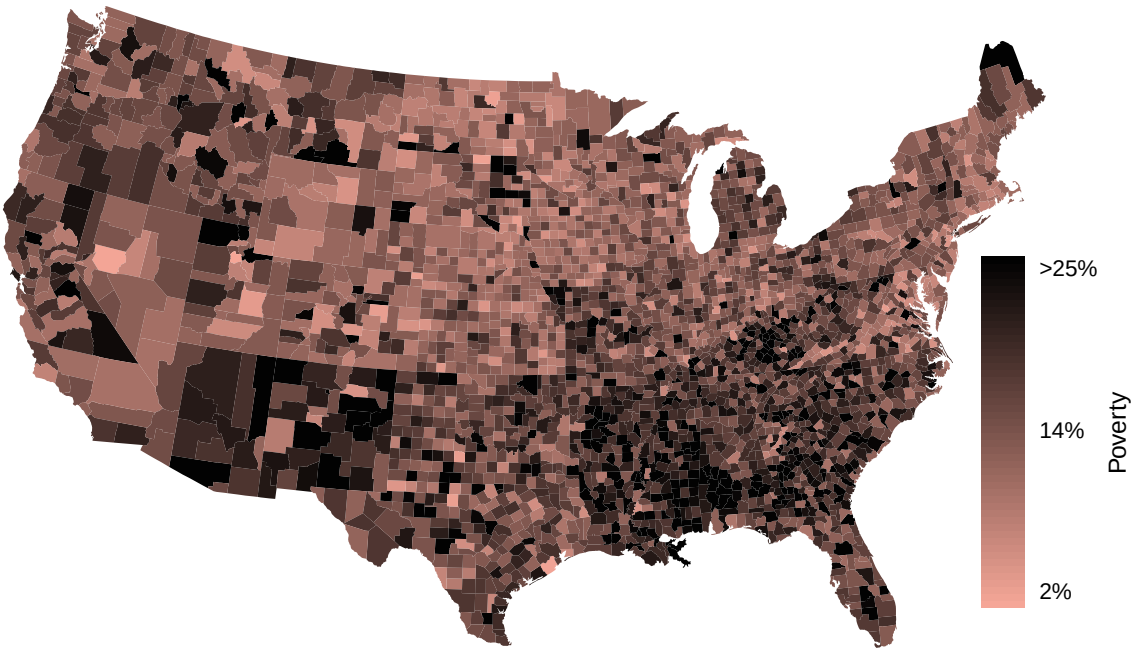
Poverty rates are evidently higher in a few locations. Notably, the deep south shows higher poverty rates, as does much of Arizona and New Mexico. High poverty rates are evident in the Mississippi flood plains a little north of New Orleans and also in a large section of Kentucky.

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

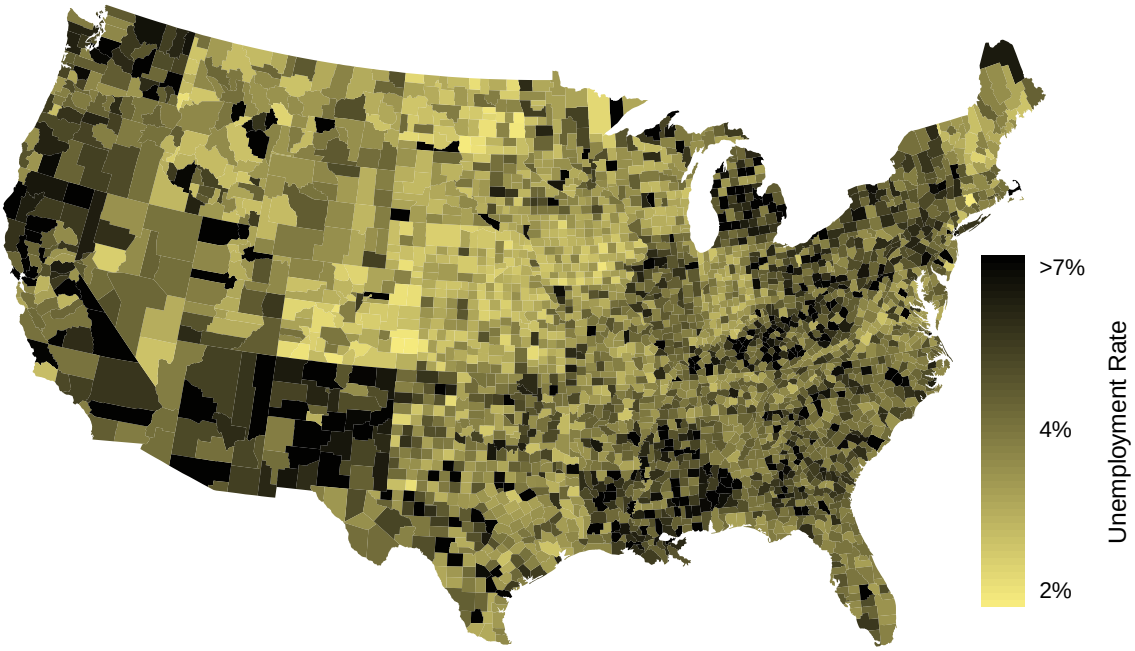
GUIDED PRACTICE 2.22

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

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

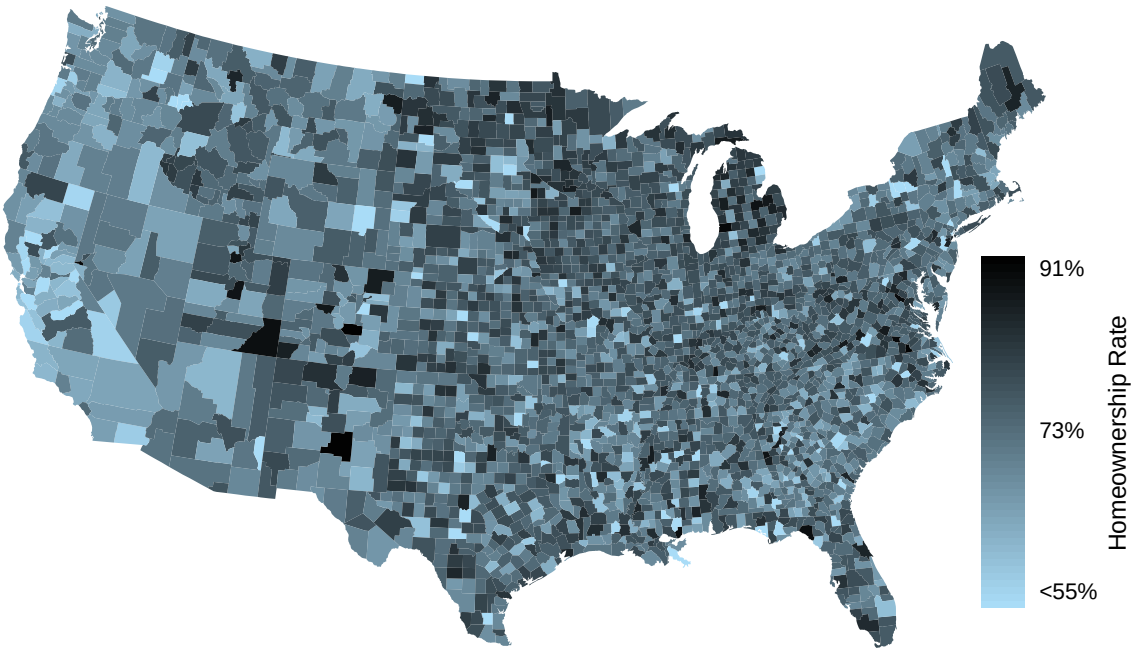


(a)

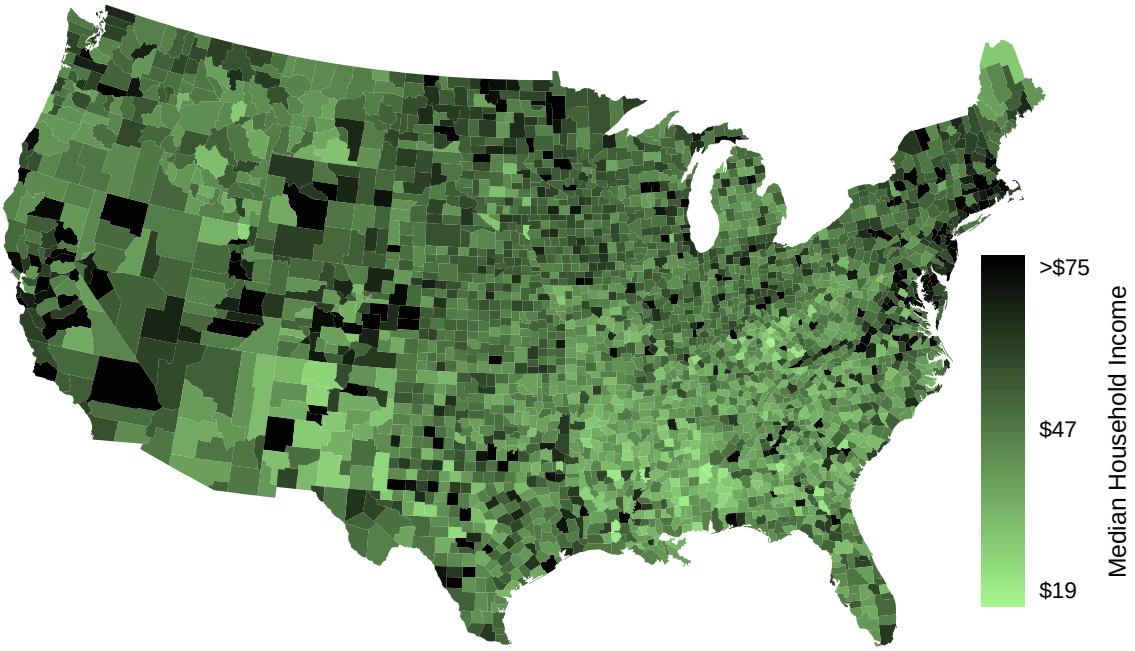


(b)

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



(a)



(b)

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