

Chapter 2

Summarizing Data

After collecting data, the next stage in the investigative process is to summarize the data. Graphical displays allow us to visualize and better understand the important features of a data set.

2.1 Examining numerical data

2.1.2 Stem-and-leaf plots and dot plots

Sometimes two variables is one too many: only one variable may be of interest. In these cases we want to focus not on the association between two variables, but on the distribution of a single variable. The term **distribution** refers to the values that a variable takes and the frequency of these values. Let's take a closer look at the `email50` data set and focus on the number of characters in each email. To simplify the data, we will round the numbers and record the values in thousands. Thus, 22105 is recorded as 22.

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

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

Rather than look at the data as a list of numbers, which makes the distribution difficult to discern, we will organize it into a table called a **stem-and-leaf plot** shown in Figure 2.4. In a stem-and-leaf plot, each number is broken into two parts. The first part is called the **stem** and consists of the beginning digit(s). The second part is called the **leaf** and consists of the final digit(s). The stems are written in a column in ascending order, and the leaves that match up with those stems are written on the corresponding row. Figure 2.4 shows a stem-and-leaf plot of the number of characters in 50 emails. The stem represents the ten thousands place and the leaf represents the thousands place. For example, 1 | 2 corresponds to 12 thousand. When making a stem-and-leaf plot, remember

⁴Consider 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 since water becomes toxic when consumed in excessive quantities.

to include a legend that describes what the stem and what the leaf represent. Without this, there is no way of knowing if $1 \mid 2$ represents 1.2, 12, 120, 1200, etc.

```

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

```

Legend: $1 \mid 2 = 12,000$

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

- ⊙ **Guided Practice 2.6** There are a lot of numbers on the first row of the stem-and-leaf plot. Why is this the case?⁵

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

```

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

```

Legend: $1 \mid 2 = 12,000$

Figure 2.5: A split stem-and-leaf.

- ⊙ **Guided Practice 2.7** What is the smallest number in this data set? What is the largest?⁶

⁵There are a lot of numbers on the first row because there are a lot of values in the data set less than 10 thousand.

⁶The smallest number is less than 1 thousand, and the largest is 64 thousand. That is a big range!

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

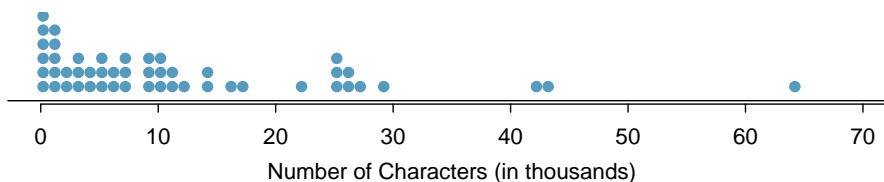


Figure 2.6: A dot plot of `num_char` for the `email50` data set.

- **Guided Practice 2.8** Imagine rotating the dot plot 90 degrees clockwise. What do you notice?⁷

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

- **Example 2.9** Based on both the stem-and-leaf and dot plot, where are the values clustered and where are the gaps for the `email50` data set?

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

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

Outliers are extreme

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

TIP: Why it is important to look for outliers

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

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

⁷It has a similar shape as the stem-and-leaf plot! The values on the horizontal axis correspond to the stems and the number of dots in each interval correspond the number of leaves needed for each stem.

- ⊙ **Guided Practice 2.10** The observation 64 thousand, a suspected outlier, was found to be an accurate observation. What would such an observation suggest about the nature of character counts in emails?⁸
- ⊙ **Guided Practice 2.11** Consider a data set that consists of the following numbers: 12, 12, 12, 12, 12, 13, 13, 14, 14, 15, 19. Which graph would better illustrate the data: a stem-and-leaf plot or a dot plot? Explain.⁹

2.1.3 Histograms and shape

Stem-and-leaf plots and dot plots are ideal for displaying data from small samples because they show the exact values of the observations and how frequently they occur. However, they are impractical for larger samples. Rather than showing the frequency of every value, we prefer to think of the value as belonging to a *bin*. For example, in the `email150` data set, we create a table of counts for the number of cases with character counts between 0 and 5,000, then the number of cases between 5,000 and 10,000, and so on. Such a table, shown in Table 2.7, is called a **frequency table**. Observations that fall on the boundary of a bin (e.g. 5,000) are allocated to the lower bin.¹⁰ These binned counts are plotted as bars in Figure 2.9 into what is called a **histogram** or **frequency histogram**, which resembles the stacked dot plot shown in Figure 2.6.

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

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

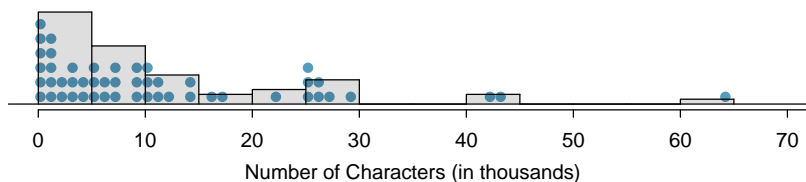


Figure 2.8: A histogram of `num_char`. This histogram is drawn over the corresponding dot plot.

TIP: Drawing histograms

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

⁸That occasionally there may be very long emails.

⁹Because all the values begin with 1, there would be only one stem (or two in a split stem-and-leaf). This would not provide a good sense of the distribution. For example, the gap between 15 and 19 would not be visually apparent. A dot plot would be better here.

¹⁰This is called *left inclusive*.

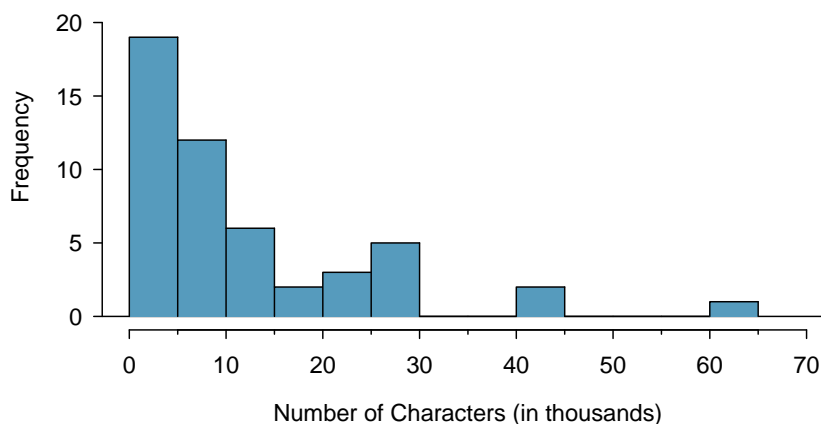


Figure 2.9: A histogram of `num_char`. This histogram uses bins or class intervals of width 5.

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

- **Example 2.12** How many emails had fewer than 10 thousand characters?

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

- **Example 2.13** Approximately how many emails had fewer than 1 thousand characters?

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

- **Example 2.14** What *percent* of the emails had 10 thousand or more characters?

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

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

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

Table 2.10: The cumulative frequencies for the binned `num_char` data.

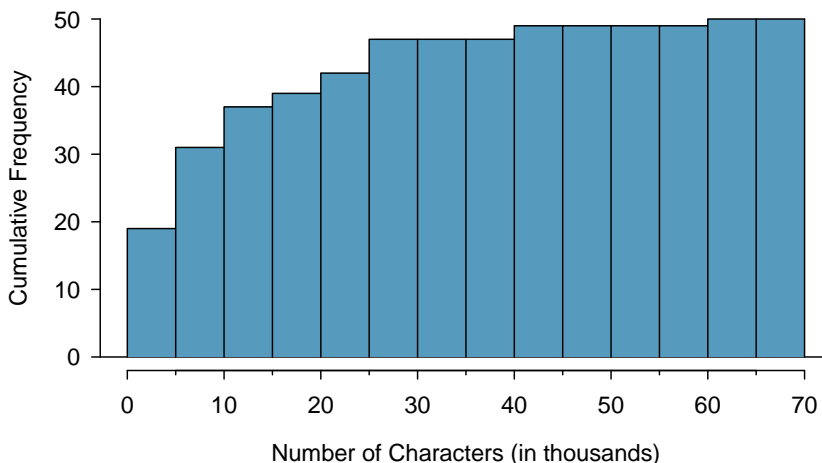


Figure 2.11: A cumulative frequency histogram of `num_char`. This histogram uses bins or class intervals of width 5.

- **Example 2.15** How many of the emails had fewer than 20 thousand characters?

By tracing the height of the 15-20 thousand bin over to the vertical axis, we can see that it has a height just under 40 on the cumulative frequency scale. Therefore, we estimate that ≈ 39 of the emails had fewer than 30 thousand characters. Note that answering this question using the original frequency histogram would require additional work.

- **Example 2.16** Using the cumulative frequency histogram, how many of the emails had 10-15 thousand characters?

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

- **Example 2.17** Approximately 25 of the emails had fewer than how many characters?

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

Knowing that 25 of the emails had fewer than a value between 5 and 10 thousand characters is useful information, but it is even more useful if we know what percent of the total 25 represents. Knowing that there were 50 total emails tells us that $25/50 = 0.5 = 50\%$ of the emails had fewer than a value between 5 and 10 thousand characters.

When we want to know what fraction or percent of the data meet a certain criteria, we use relative frequency instead of frequency. **Relative frequency** is a fancy term for percent or proportion. It tells us how large a number is relative to the total.

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

- ⦿ **Guided Practice 2.18** How will the *shape* of the relative frequency histograms differ from the frequency histograms?¹¹

Caution: Pay close attention to the vertical axis of a histogram

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

Frequency and relative frequency histograms are especially convenient for describing the **shape** of the data distribution. Figure 2.9 shows that most emails have a relatively small number of characters, while fewer emails have a very large number of characters. When data trail off to the right in this way and have a longer right tail, the shape is said to be **right skewed**.¹²

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

Long tails to identify skew

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

- ⦿ **Guided Practice 2.19** Take a look at the dot plot in Figure 2.6. Can you see the skew in the data? Is it easier to see the skew in the frequency histogram, the dot plot, or the stem-and-leaf plot?¹³
- ⦿ **Guided Practice 2.20** What can you see in the dot plot and stem-and-leaf plot that you cannot see in the frequency histogram?¹⁴

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.¹⁵ There is only one prominent peak in the histogram of `num_char`.

¹¹The shape will remain exactly the same. Changing from frequency to relative frequency involves dividing all the frequencies by the same number, so only the vertical scale (the numbers on the y-axis) change.

¹²Other ways to describe data that are skewed to the right: **skewed to the right**, **skewed to the high end**, or **skewed to the positive end**.

¹³The skew is visible in all three plots. It is not easily visible in the cumulative frequency histogram.

¹⁴Character counts for individual emails.

¹⁵Another definition of mode, which is not typically used in statistics, is the value with the most occurrences. It is common to have *no* observations with the same value in a data set, which makes this other definition useless for many real data sets.

Figure 2.12 shows histograms that have one, two, or three prominent peaks. Such distributions are called **unimodal**, **bimodal**, and **multimodal**, respectively. Any distribution with more than 2 prominent peaks is called multimodal. Notice that in Figure 2.9 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.

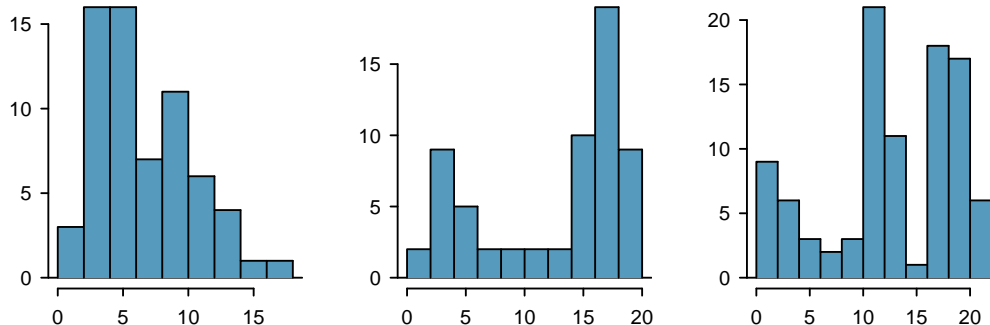


Figure 2.12: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

- ⊙ **Guided Practice 2.21** Height measurements of young students and adult teachers at a K-3 elementary school were taken. How many modes would you anticipate in this height data set?¹⁶

TIP: Looking for modes

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

2.2 Numerical summaries and box plots

2.2.1 Measures of center

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

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

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.6 \quad (2.22)$$

¹⁶There 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.

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

\bar{x}
sample
mean

Mean

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

$$\bar{x} = \frac{1}{n} \sum x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad (2.23)$$

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

- ⊙ **Guided Practice 2.24** Examine Equations (2.22) and (2.23) above. What does x_1 correspond to? And x_2 ? What does x_i represent?¹⁷
- ⊙ **Guided Practice 2.25** What was n in this sample of emails?¹⁸

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

μ
population
mean

- **Example 2.26** The average number of characters across all emails can be estimated using the sample data. Based on the sample of 50 emails, what would be a reasonable estimate of μ_x , the mean number of characters in all emails in the `email` data set? (Recall that `email50` is a sample from `email`.)

The sample mean, 11,600, may provide a reasonable estimate of μ_x . While this number will not be perfect, it provides a *point estimate* of the population mean. In Chapter 5 and beyond, we will develop tools to characterize the reliability of point estimates, and we will find that point estimates based on larger samples tend to be more reliable than those based on smaller samples.

¹⁷ x_1 corresponds to the number of characters in the first email in the sample (21.7, in thousands), x_2 to the number of characters in the second email (7.0, in thousands), and x_i corresponds to the number of characters in the i^{th} email in the data set.

¹⁸The sample size was $n = 50$.

- **Example 2.27** We might like to compute the average income per person in the US. To do so, we might first think to take the mean of the per capita incomes across the 3,143 counties in the `county` data set. What would be a better approach?

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

Example 2.27 used what is called a **weighted mean**, which will not be a key topic in this textbook. However, we have provided an online supplement on weighted means for interested readers:

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

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

Median: the number in the middle

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

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

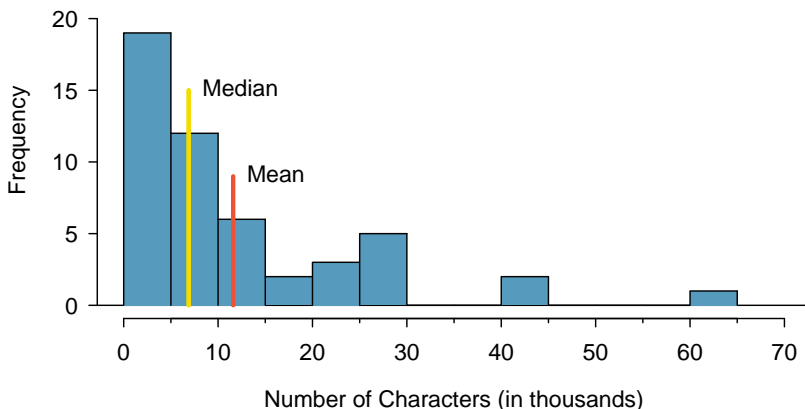


Figure 2.13: A histogram of `num_char` with its mean and median shown.

- **Example 2.28** Based on the data, why is the mean greater than the median in this data set?

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

The mean follows the tail

In a right skewed distribution, the mean is greater than the median.

In a left skewed distribution, the mean is less than the median.

In a symmetric distribution, the mean and median are approximately equal.

- **Guided Practice 2.29** Consider the distribution of individual income in the United States. Which is greater: the mean or median? Why? ¹⁹

2.2.2 Standard deviation as a measure of spread

The U.S. Census Bureau reported that in 2012, the median family income was \$62,241 and the mean family income was \$82,743.²⁰

Is a family income of \$40,000 relatively low, or is it a somewhat typical income? In order to answer this question, it is not enough to know the center of the data set and its **range** (maximum value - minimum value). We must know about the variability of the data set within that range. Low variability or small spread means that the values tend to be more clustered together. High variability or large spread means that the values tend to be far apart.

- **Example 2.30** Is it possible for two data sets to have the same range but different spread? If so, give an example. If not, explain why not.

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

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

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

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

$$x_1 - \bar{x} = 21.7 - 11.6 = 10.1$$

$$x_2 - \bar{x} = 7.0 - 11.6 = -4.6$$

$$x_3 - \bar{x} = 0.6 - 11.6 = -11.0$$

$$\vdots$$

$$x_{50} - \bar{x} = 15.8 - 11.6 = 4.2$$

¹⁹Because a small percent of individuals earn extremely large amounts of money while the majority earn a modest amount, the distribution is skewed to the right. Therefore, the mean is greater than the median.

²⁰<http://www.census.gov/hhes/www/income/>

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

s^2
sample
variance

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

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

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

s
sample
standard
deviation

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

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

Calculating the Standard Deviation

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

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} \quad (2.31)$$

The variance is useful for mathematical reasons, but the standard deviation is easier to interpret because it has the same units as the data set. The units for variance will be the units squared (e.g. meters²). Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.²¹ However, like the mean, the population values have special symbols: σ^2 for the variance and σ for the standard deviation. The symbol σ is the Greek letter *sigma*.

σ^2
population
variance

TIP: standard deviation describes variability

Focus on the conceptual meaning of the standard deviation as a descriptor of variability rather than the formulas. The empirical rule tells us that usually about 68% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations of the mean. However, as seen in Figures 2.14 and 2.15, these percentages are not strict rules.²²

σ
population
standard
deviation

²¹The only difference is that the population variance has a division by n instead of $n - 1$.

²²We will learn where these two numbers come from in Chapter 4 when we study the normal distribution.

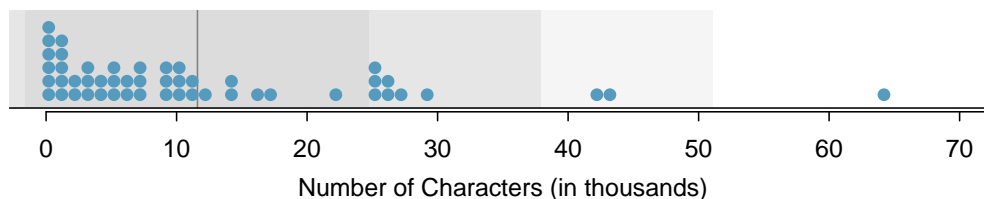


Figure 2.14: In the `num_char` data, 40 of the 50 emails (80%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 68% (or approximately $2/3$) of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule of thumb is less accurate for skewed data, as shown in this example.

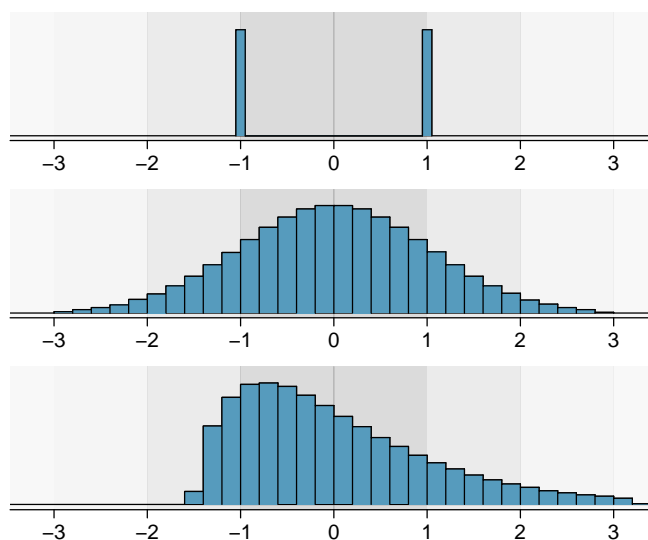


Figure 2.15: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

- ⊙ **Guided Practice 2.32** On page 43, 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 2.15 as an example, explain why such a description is important.²³

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

²³Figure 2.15 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.

- **Example 2.33** In the data’s context (the number of characters in emails), describe the distribution of the `num_char` variable using the histogram in Figure 2.16.

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

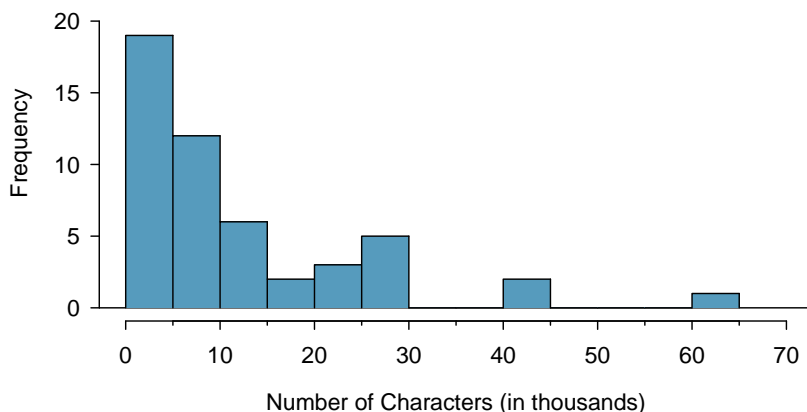


Figure 2.16: A copy of Figure 2.9.

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

2.2.3 Box plots and quartiles

A **box plot** summarizes a data set using five summary statistics while also plotting unusual observations. Figure 2.17 provides a box plot of the `num_char` variable from the `email50` data set.

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

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

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

To determine if there are any unusually distant observations (i.e. outliers), we first calculate the spread of the middle 50% of the data by subtracting Q_1 from Q_3 : $Q_3 - Q_1 = 13,375$. This quantity is called the **interquartile range** (IQR, for short). It, like the standard deviation, is a measure of variability or **spread** in data. The more variable the data, the larger the standard deviation and IQR tend to be.

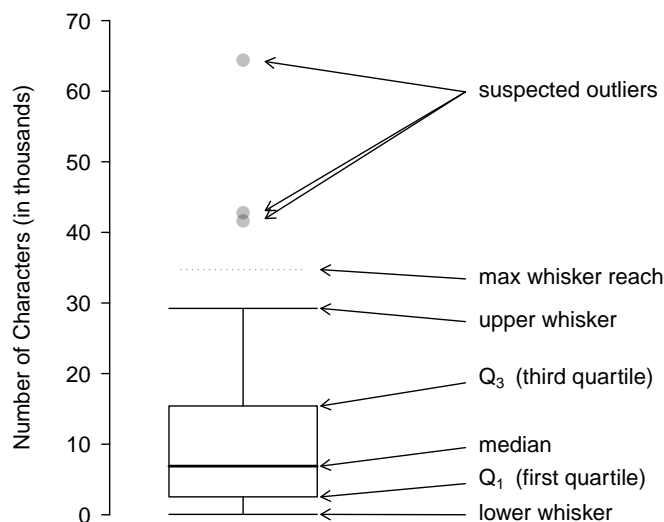


Figure 2.17: A labeled box plot for the number of characters in 50 emails. The median (6,890) splits the data into the bottom 50% and the top 50%.

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.

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

Extending out from the rectangle, the **whiskers** attempt to capture the data remaining outside of the box; however, their reach cannot be more than $1.5 \times IQR$. In Figure 2.17, the upper whisker does not extend to the last three points, which are beyond $Q_3 + 1.5 \times IQR$, and so it extends only to the last point below this limit. The lower whisker stops at the lowest value, 33, since there is no additional data to reach. 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. We will call a value an outlier if it is more than $1.5 \times IQR$ below Q_1 or above Q_3 .²⁴

- **Example 2.34** Compare the box plot to the graphs previously discussed: stem-and-leaf plot, dot plot, frequency and relative frequency histogram. What can we learn more easily from a box plot? What can we learn more easily from the other graphs?

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

²⁴While the choice of exactly 1.5 is arbitrary, it is the most commonly used value for box plots.

- **Example 2.35** Is it possible to identify skew from the box plot?

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

- **Guided Practice 2.36** What percent of the data fall between Q_1 and the median? What percent is between the median and Q_3 ?²⁵

2.2.4 Calculator: summarize 1-variable statistics

TI calculator: Entering data

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

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

TI calculator: Calculating Summary Statistics

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

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

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

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

\bar{x}	(mean)	n	(sample size or # of data points)
Σx	(sum of all the data values)	$\min X$	(minimum)
Σx^2	(sum of all the squared data values)	Q_1	(first quartile)
S_x	(sample standard deviation)	Med	(median)
σ_x	(population standard deviation)	Q_3	(third quartile)

²⁵ Since Q_1 is the 25th percentile and Q_2 is the 50th percentile, 25% of the data falls between Q_1 and Q_2 . Similarly, 25% of the data fall between Q_2 and Q_3 .

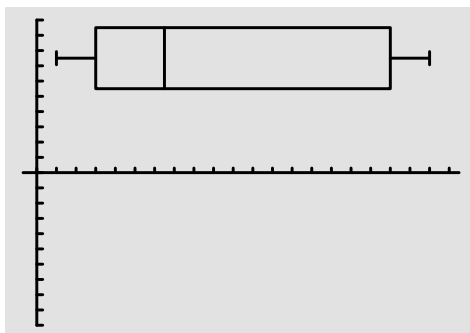
TI calculator: Drawing a box plot

Occasionally, we may want just a quick sketch of the box plot. In these instances we can use the graphing calculator to speed up the process. Use **2ND Y=**.

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

- **Example 2.37** Enter the following 10 data points into list L1 on a calculator: 5, 8, 1, 19, 3, 1, 11, 18, 20, 5. Find the summary statistics and make a box plot of the data.

The summary statistics should be $\bar{x}=9.1$, $S_x = 7.475$, $Q_1 = 3$, etc. The box plot should be as follows.

**TI calculator: What to do if you cannot find L1 or another list**

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

Lists L1 - L6 will now be restored.

2.2.5 Outliers and robust statistics

Rules of thumb for identifying outliers

There are two rules of thumb for identifying outliers:

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

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

- ⊙ **Guided Practice 2.38** For the `email150` data set, $Q_1 = 2,536$ and $Q_3 = 15,411$. $\bar{x} = 11,600$ and $s = 13,130$. What values would be considered an outlier on the low end using each rule?²⁶
- ⊙ **Guided Practice 2.39** Because there are no negative values in this data set, there can be no outliers on the low end. What does the fact that there are outliers on the high end but not on the low end suggest?²⁷

How are the sample statistics of the `num_char` data set affected by the observation, 64,401? What would have happened if this email wasn't observed? What would happen to these summary statistics if the observation at 64,401 had been even larger, say 150,000? These scenarios are plotted alongside the original data in Figure 2.18, and sample statistics are computed under each scenario in Table 2.19.

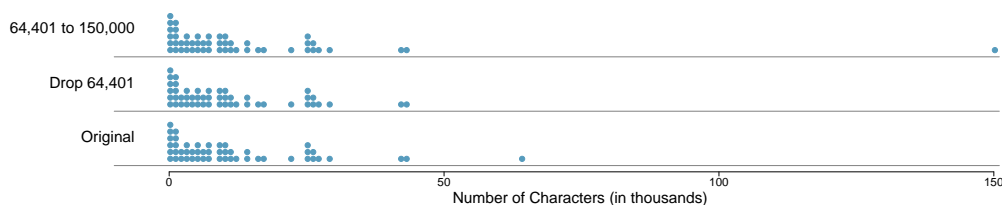


Figure 2.18: Dot plots of the original character count data and two modified data sets.

- ⊙ **Guided Practice 2.40** (a) Which is more affected by extreme observations, the mean or median? Table 2.19 may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?²⁸

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

²⁶ $Q_1 - 1.5 \times IQR = 2536 - 1.5 \times (15411 - 2536) = -16,749.5$, so values less than $-16,749.5$ would be considered an outlier using the first rule of thumb. Using the second rule of thumb, a value less than $\bar{x} - 2 \times s = 11,600 - 2 \times 13,130 = -14,660$ would be considered an outlier. Note that these are just rules of thumb and yield different values.

²⁷It suggests that the distribution has a right hand tail, that is, that it is right skewed.

²⁸(a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Guided Practice 2.40.

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

Table 2.19: A comparison of how the median, IQR, mean (\bar{x}), and standard deviation (s) change when extreme observations are present.

- **Example 2.41** The median and IQR do not change much under the three scenarios in Table 2.19. Why might this be the case?

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

- **Guided Practice 2.42** The distribution of vehicle prices tends to be right skewed, with a few luxury and sports cars lingering out into the right tail. If you were searching for a new car and cared about price, should you be more interested in the mean or median price of vehicles sold, assuming you are in the market for a regular car?²⁹

2.2.6 Linear transformations of data

- **Example 2.43** Begin with the following list: 1, 1, 5, 5. Multiply all of the numbers by 10. What happens to the mean? What happens to the standard deviation? How do these compare to the mean and the standard deviation of the original list?

The original list has a mean of 3 and a standard deviation of 2. The new list: 10, 10, 50, 50 has a mean of 30 with a standard deviation of 20. Because all of the values were multiplied by 10, both the mean and the standard deviation were multiplied by 10.³⁰

- **Example 2.44** Start with the following list: 1, 1, 5, 5. Multiply all of the numbers by -0.5. What happens to the mean? What happens to the standard deviation? How do these compare to the mean and the standard deviation of the original list?

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

²⁹Buyers of a “regular car” should be concerned about the median price. High-end car sales can drastically inflate the mean price while the median will be more robust to the influence of those sales.

³⁰Here, the population standard deviation was used in the calculation. These properties can be proven mathematically using properties of sigma (summation).

- **Example 2.45** Again, start with the following list: 1, 1, 5, 5. Add 100 to every entry. How do the new mean and standard deviation compare to the original mean and standard deviation?

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

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

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

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

$$aX + b$$

Using the examples as a guide, we can solve this temperature-conversion problem. The mean was 27°C and the standard deviation was 3°C . To convert to Fahrenheit, we multiply all of the values by $9/5$, which multiplies both the mean and the standard deviation by $9/5$. Then we add 32 to all of the values which adds 32 to the mean but does not change the standard deviation further.

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

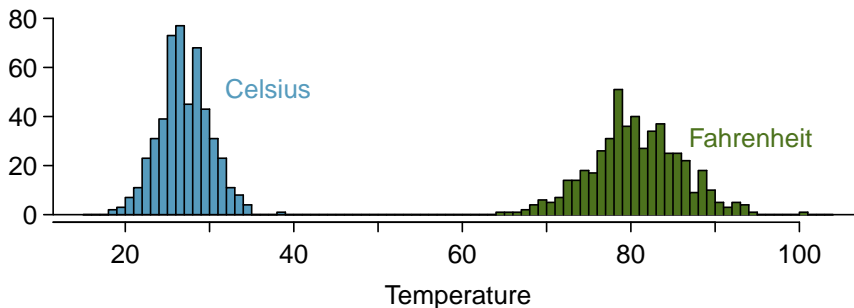


Figure 2.20: 500 temperatures shown in both Celsius and Fahrenheit.

Adding shifts the values, multiplying stretches or contracts them

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

- **Example 2.46** Consider the temperature example. How would converting from Celsius to Fahrenheit affect the median? The IQR?

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

For a more mathematical explanation of the IQR calculation, see the footnote.³¹

2.2.7 Comparing numerical data across groups

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

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

There were 2,041 counties where the population increased from 2000 to 2010, and there were 1,099 counties with no gain (all but one were a loss). A random sample of 100 counties from the first group and 50 from the second group are shown in Table 2.21 to give a better sense of some of the raw data, and Figure 2.22 shows a **back-to-back stem-and-leaf plot**.

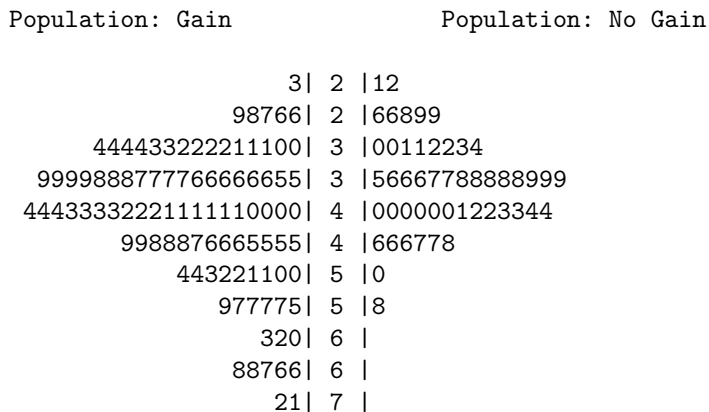
The **parallel box plot** is a traditional tool for comparing across groups. An example is shown in the left panel of Figure 2.23, where there are two box plots, one for each group, placed into one plotting window and drawn on the same scale.

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 2.23.

³¹ $\text{new IQR} = \left(\frac{9}{5}Q_3 + 32\right) - \left(\frac{9}{5}Q_1 + 32\right) = \frac{9}{5}(Q_3 - Q_1) = \frac{9}{5} \times (\text{old IQR}).$

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

Table 2.21: In this table, median household income (in \$1000s) from a random sample of 100 counties that gained population over 2000-2010 are shown on the left. Median incomes from a random sample of 50 counties that had no population gain are shown on the right.



Legend: 4 | 5 = 45,000 median income

Figure 2.22: Back-to-back stem-and-leaf plot for median income, split by whether the county had a population gain or no gain.

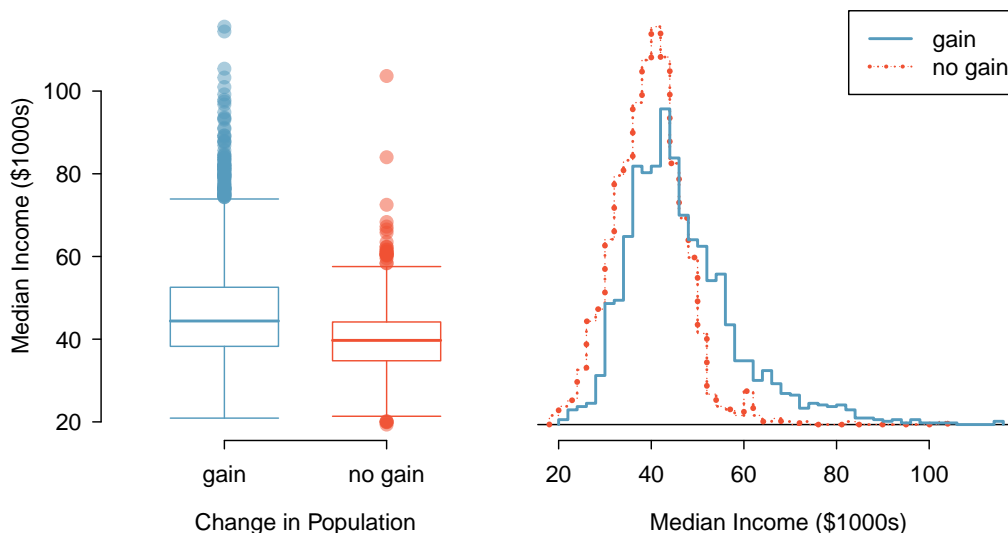


Figure 2.23: Side-by-side box plot (left panel) and hollow histograms (right panel) for `med_income`, where the counties are split by whether there was a population gain or loss from 2000 to 2010. The income data were collected between 2006 and 2010.

- ⦿ **Guided Practice 2.47** Use the plots in Figure 2.23 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?³²

TIP: Comparing distributions

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

- ⦿ **Guided Practice 2.48** What components of each plot in Figure 2.23 do you find most useful?³³
- ⦿ **Guided Practice 2.49** Do these graphs tell us about any association between income for the two groups?³⁴

³²Answers may vary a little. The counties with population gains tend to have higher income (median of about \$45,000) versus counties without a gain (median of about \$40,000). The variability is also slightly larger for the population gain group. This is evident in the IQR, which is about 50% bigger in the *gain* group. Both distributions show slight to moderate right skew and are unimodal. There is a secondary small bump at about \$60,000 for the *no gain* group, visible in the hollow histogram plot, that seems out of place. (Looking into the data set, we would find that 8 of these 15 counties are in Alaska and Texas.) The box plots indicate there are many observations far above the median in each group, though we should anticipate that many observations will fall beyond the whiskers when using such a large data set.

³³Answers will vary. The parallel box plots are especially useful for comparing centers and spreads, while the hollow histograms are more useful for seeing distribution shape, skew, and groups of anomalies.

³⁴No, to see association we require a scatterplot. Moreover, these data are not paired, so the discussion of association does not make sense here.

2.2.8 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 map it using an **intensity map**, where colors are used to show higher and lower values of a variable. Figures 2.24 and 2.25 shows intensity maps for federal spending per capita (`fed_spend`), poverty rate in percent (`poverty`), homeownership rate in percent (`homeownership`), and median household income (`med_income`). The color key indicates which colors correspond to which values. Note that the intensity maps are not generally very helpful for getting precise values in any given county, but they are very helpful for seeing geographic trends and generating interesting research questions.

- **Example 2.50** What interesting features are evident in the `fed_spend` and `poverty` intensity maps?

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

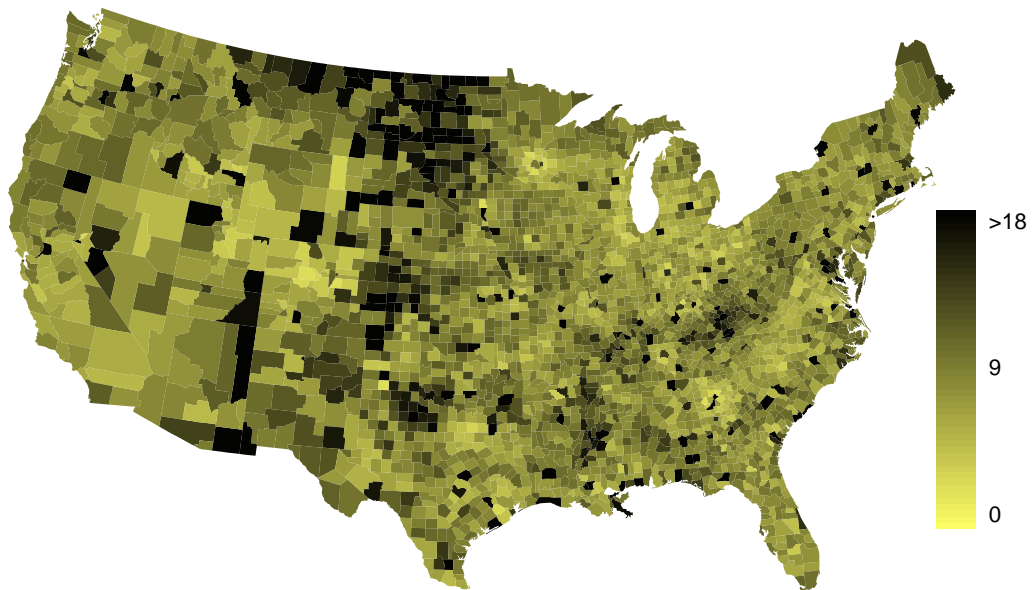
Poverty rates are evidently higher in a few locations. Notably, the deep south shows higher poverty rates, as does the southwest border of Texas. The vertical strip of eastern Utah and Arizona, noted above for its higher federal spending, also appears to have higher rates of poverty (though generally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

- **Guided Practice 2.51** What interesting features are evident in the `med_income` intensity map?³⁵

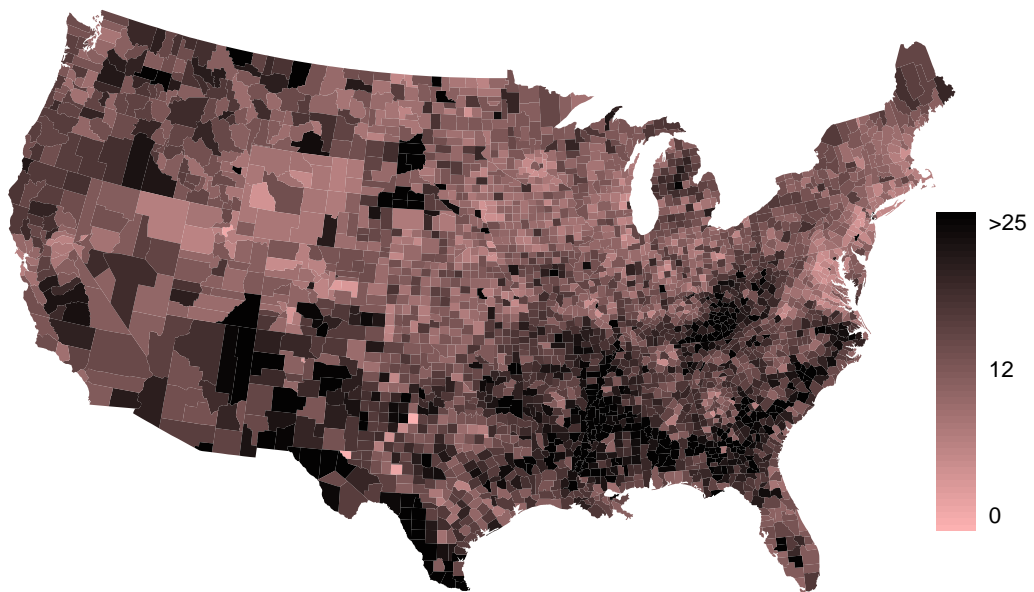
2.3 Considering categorical data

Like numerical data, categorical data can also be organized and analyzed. In this section, we will introduce tables and other basic tools for categorical data that are used throughout this book. The `email150` data set represents a sample from a larger email data set called `email`. This larger data set contains information on 3,921 emails. In this section we will examine whether the presence of numbers, small or large, in an email provides any useful value in classifying email as spam or not spam.

³⁵Note: answers will vary. There is a very strong correspondence between high earning and metropolitan areas. You might look for large cities you are familiar with and try to spot them on the map as dark spots.

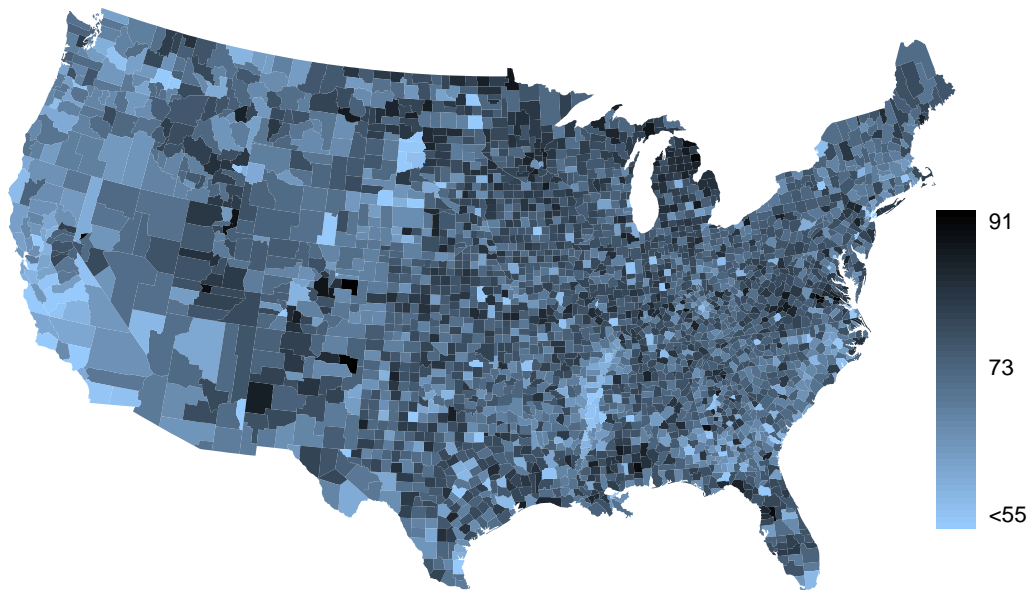


(a)

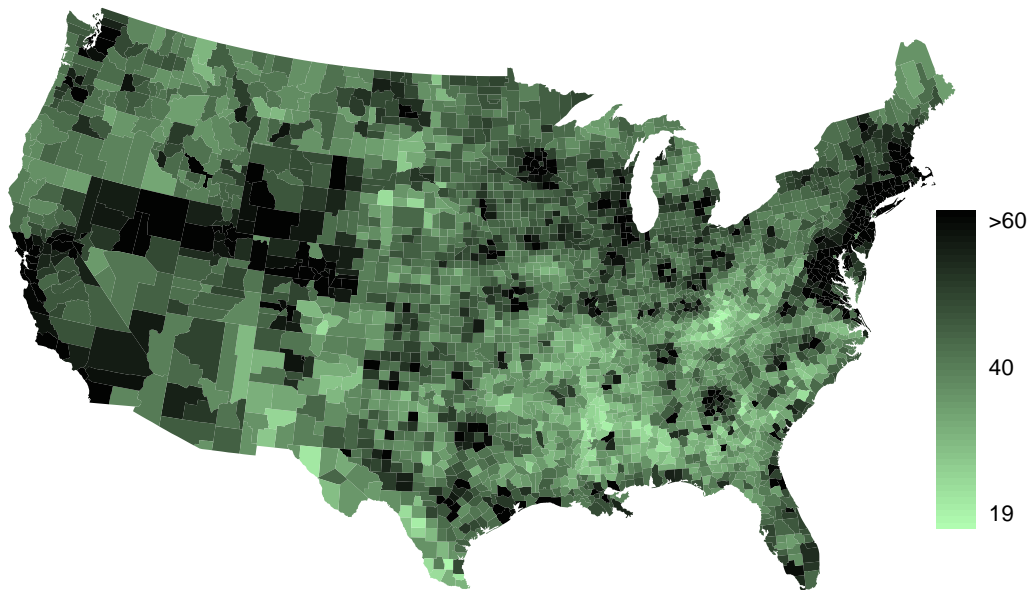


(b)

Figure 2.24: (a) Map of federal spending (dollars per capita). (b) Intensity map of poverty rate (percent).



(a)



(b)

Figure 2.25: (a) Intensity map of homeownership rate (percent). (b) Intensity map of median household income (\$1000s).

2.3.1 Contingency tables and bar plots

Table 2.26 summarizes two variables: **spam** and **number**. Recall that **number** is a categorical variable that describes whether an email contains no numbers, only small numbers (values under 1 million), or at least one big number (a value of 1 million or more). A table that summarizes data for two categorical variables in this way is called a **contingency table**. Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, the value 149 corresponds to the number of emails in the data set that are spam *and* had no number listed in the email. Row and column totals are also included. The **row totals** provide the total counts across each row (e.g. $149 + 168 + 50 = 367$), and **column totals** are total counts down each column.

Table 2.27 shows a frequency table for the **number** variable. If we replaced the counts with percentages or proportions, the table is a **relative frequency table**.

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

Table 2.26: A contingency table for **spam** and **number**.

none	small	big	Total
549	2827	545	3921

Table 2.27: A frequency table for the **number** variable.

Because the numbers in these tables are counts, not to data points, they cannot be graphed using the methods we applied to numerical data. Instead, another set of graphing methods are needed that are suitable for categorical data.

A bar plot is a common way to display a single categorical variable. The left panel of Figure 2.28 shows a **bar plot** for the **number** variable. In the right panel, the counts are converted into proportions (e.g. $549/3921 = 0.140$ for **none**), showing the proportion of observations that are in each level (i.e. in each category).

2.3.2 Row and column proportions

Table 2.29 shows the row proportions for Table 2.26. The **row proportions** are computed as the counts divided by their row totals. The value 149 at the intersection of **spam** and **none** is replaced by $149/367 = 0.406$, i.e. 149 divided by its row total, 367. So what does 0.406 represent? It corresponds to the proportion of spam emails in the sample that do not have any numbers.

A contingency table of the column proportions is computed in a similar way, where each **column proportion** is computed as the count divided by the corresponding column total. Table 2.30 shows such a table, and here the value 0.271 indicates that 27.1% of emails with no numbers were spam. This rate of spam is much higher compared to emails with only small numbers (5.9%) or big numbers (9.2%). Because these spam rates vary between the three levels of **number** (**none**, **small**, **big**), this provides evidence that the **spam** and **number** variables are associated.

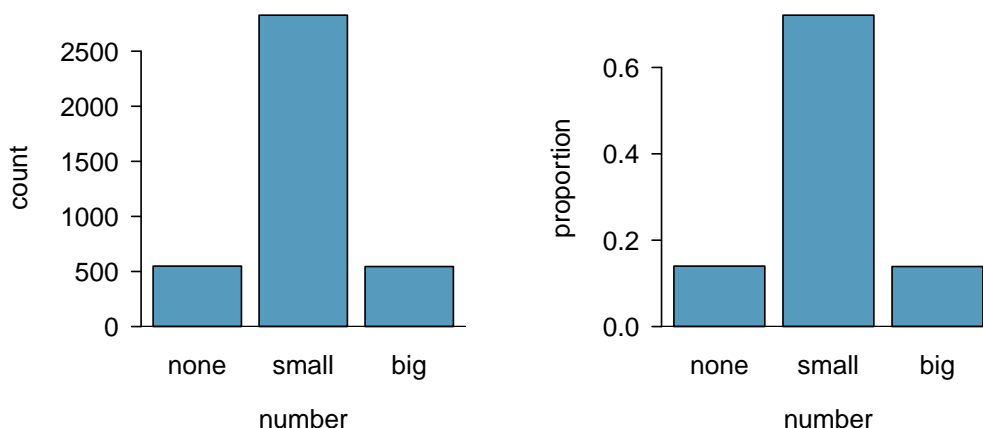


Figure 2.28: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

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

Table 2.29: A contingency table with row proportions for the `spam` and `number` variables.

We could also have checked for an association between `spam` and `number` in Table 2.29 using row proportions. When comparing these row proportions, we would look down columns to see if the fraction of emails with no numbers, small numbers, and big numbers varied from `spam` to `not spam`.

- ⦿ **Guided Practice 2.52** What does 0.458 represent in Table 2.29? What does 0.059 represent in Table 2.30?³⁶
- ⦿ **Guided Practice 2.53** What does 0.139 at the intersection of `not spam` and `big` represent in Table 2.29? What does 0.908 represent in the Table 2.30?³⁷

³⁶0.458 represents the proportion of spam emails that had a small number. 0.058 represents the fraction of emails with small numbers that are spam.

³⁷0.139 represents the fraction of non-spam email that had a big number. 0.908 represents the fraction of emails with big numbers that are non-spam emails.

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

Table 2.30: A contingency table with column proportions for the **spam** and **number** variables.

- **Example 2.54** Data scientists use statistics to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One of those characteristics is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is whether or not an email has any HTML content. A contingency table for the **spam** and **format** variables from the **email** data set are shown in Table 2.31. Recall that an HTML email is an email with the capacity for special formatting, e.g. bold text. In Table 2.31, which would be more helpful to someone hoping to classify email as spam or regular email: row or column proportions?

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

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

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

Table 2.31: A contingency table for **spam** and **format**.

Example 2.54 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.

- ⊙ **Guided Practice 2.55** Look back to Tables 2.29 and 2.30. Which would be more useful to someone hoping to identify spam emails using the **number** variable?³⁸

³⁸The column proportions in Table 2.30 will probably be most useful, which makes it easier to see that emails with small numbers are spam about 5.9% of the time (relatively rare). We would also see that about 27.1% of emails with no numbers are spam, and 9.2% of emails with big numbers are spam.

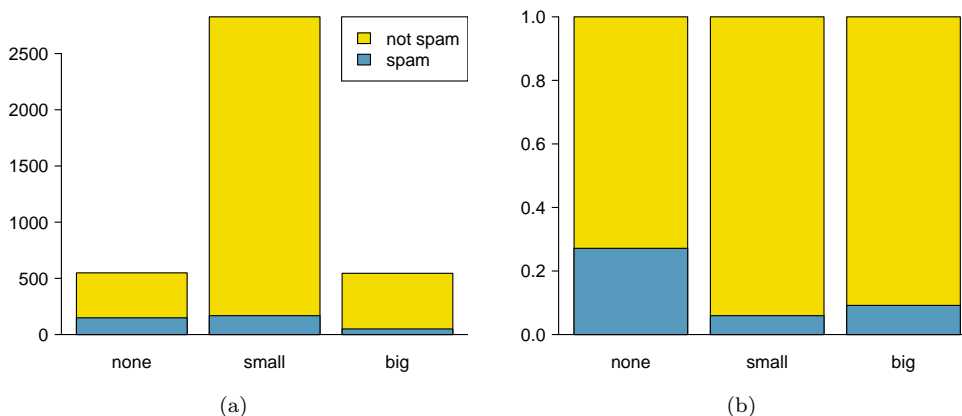


Figure 2.32: (a) Segmented bar plot for numbers found in emails, where the counts have been further broken down by `spam`. (b) Standardized version of Figure (a).

2.3.3 Segmented bar plots

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

A **segmented bar plot** is a graphical display of contingency table information. For example, a segmented bar plot representing Table 2.30 is shown in Figure 2.32(a), where we have first created a bar plot using the `number` variable and then divided each group by the levels of `spam`. The column proportions of Table 2.30 have been translated into a standardized segmented bar plot in Figure 2.32(b), which is a helpful visualization of the fraction of spam emails in each level of `number`.

- **Example 2.56** Examine both of the segmented bar plots. Which is more useful?

Figure 2.32(a) contains more information, but Figure 2.32(b) presents the information more clearly. This second plot makes it clear that emails with no number have a relatively high rate of spam email – about 27%! On the other hand, less than 10% of email with small or big numbers are spam.

Since the proportion of spam changes across the groups in Figure 2.32(b), we can conclude the variables are dependent, which is something we were also able to discern using table proportions. Because both the `none` and `big` groups have relatively few observations compared to the `small` group, the association is more difficult to see in Figure 2.32(a).

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

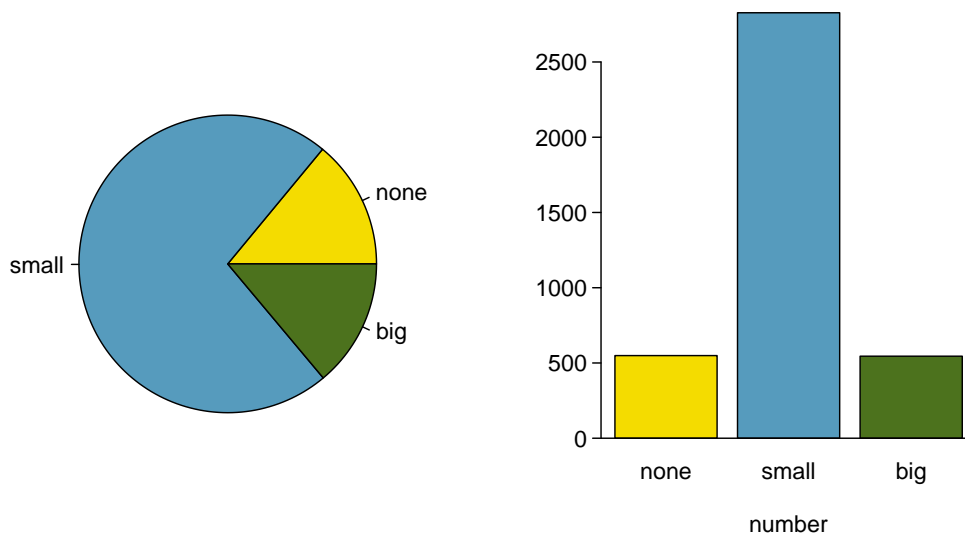


Figure 2.33: A pie chart and bar plot of `number` for the `email` data set.

2.3.4 The only pie chart you will see in this book

While pie charts are well known, they are not typically as useful as other charts in a data analysis. A **pie chart** is shown in Figure 2.33 alongside a bar plot. It is generally more difficult to compare group sizes in a pie chart than in a bar plot, especially when categories have nearly identical counts or proportions. In the case of the `none` and `big` categories, the difference is so slight you may be unable to distinguish any difference in group sizes for either plot!

Appendix A

End of chapter exercise solutions

2 Summarizing data

2.1 (a) Positive association: mammals with longer gestation periods tend to live longer as well. (b) Association would still be positive. (c) No, they are not independent. See part (a).

2.3 (a) 1/linear and 3/nonlinear. (b) 4/some curvature (nonlinearity) may be present on the right side. "Linear" would also be acceptable for the type of relationship for plot 4. (c) 2.