These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[18] were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment.

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

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

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup where doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the treatment.[19]

⊙ **Exercise 1.14** Look back to the study in Section 1.1 where researchers were testing whether stents were effective at reducing strokes in at-risk patients. Is this an experiment? Was the study blinded? Was it double-blinded?[20]

1.6 Examining numerical data

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

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

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

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

1.6.1 Scatterplots for paired data

A **scatterplot** provides a case-by-case view of data for two numerical variables. In Figure 1.8 on page 7, a scatterplot was used to examine how federal spending and poverty were related in the `county` data set. Another scatterplot is shown in Figure 1.16, comparing the number of line breaks (`line_breaks`) and number of characters (`num_char`) in emails for the `email50` data set. In any scatterplot, each point represents a single case. Since there are 50 cases in `email50`, there are 50 points in Figure 1.16.

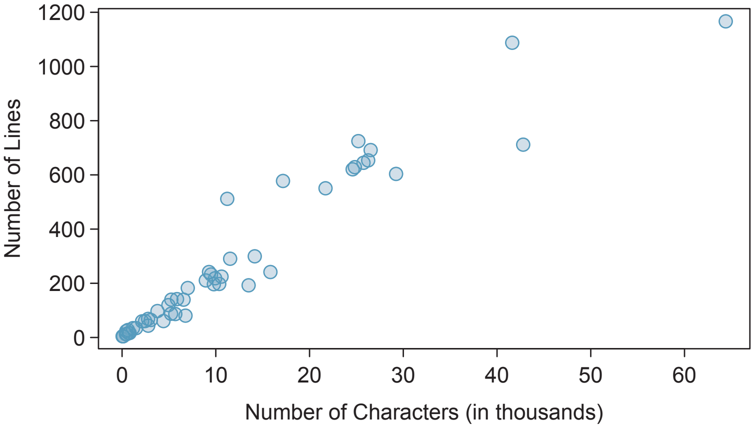

Figure 1.16: A scatterplot of `line_breaks` versus `num_char` for the `email50` data.

To put the number of characters in perspective, this paragraph has 363 characters. Looking at Figure 1.16, it seems that some emails are incredibly verbose! Upon further investigation, we would actually find that most of the long emails use the HTML format, which means most of the characters in those emails are used to format the email rather than provide text.

⊙ **Exercise 1.15** What do scatterplots reveal about the data, and how might they be useful?[21]

● **Example 1.16** Consider a new data set of 54 cars with two variables: vehicle price and weight.[22] A scatterplot of vehicle price versus weight is shown in Figure 1.17. What can be said about the relationship between these variables?

The relationship is evidently nonlinear, as highlighted by the dashed line. This is different from previous scatterplots we've seen, such as Figure 1.8 on page 7 and Figure 1.16, which show relationships that are very linear.

⊙ **Exercise 1.17** Describe two variables that would have a horseshoe shaped association in a scatterplot.[23]

[21] 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.

[22] Subset of data from http://www.amstat.org/publications/jse/v1n1/datasets.lock.html

[23] 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 since water becomes toxic when consumed in excessive quantities.

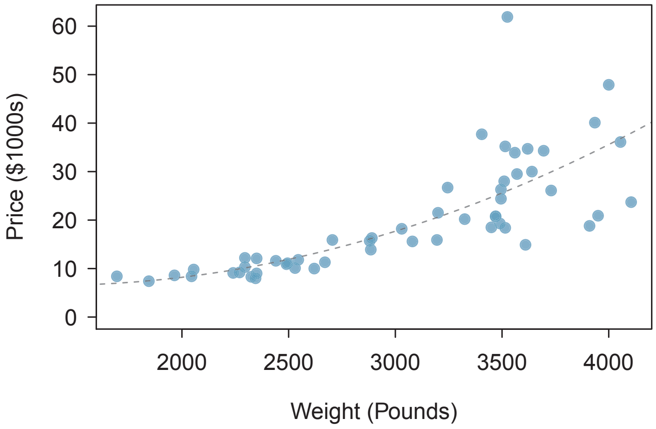

Figure 1.17: A scatterplot of `price` versus `weight` for 54 cars.

1.6.2 Dot plots and the mean

Sometimes two variables is 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 number of characters from 50 emails is shown in Figure 1.18. A stacked version of this dot plot is shown in Figure 1.19.

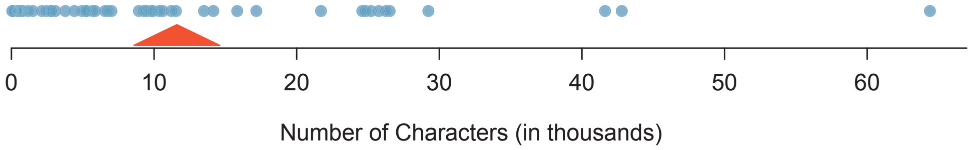

Figure 1.18: A dot plot of `num_char` for the `email50` data set.

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.

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.6 \tag{1.18}$$

The sample mean is often labeled $\bar{x}$. The letter x is being used as a generic placeholder for the variable of interest, `num_char`, and the bar over on the x communicates that the average number of characters in the 50 emails was 11,600. It is useful to think of the mean as the balancing point of the distribution. The sample mean is shown as a triangle in Figures 1.18 and 1.19.

$\bar{x}$
sample
mean

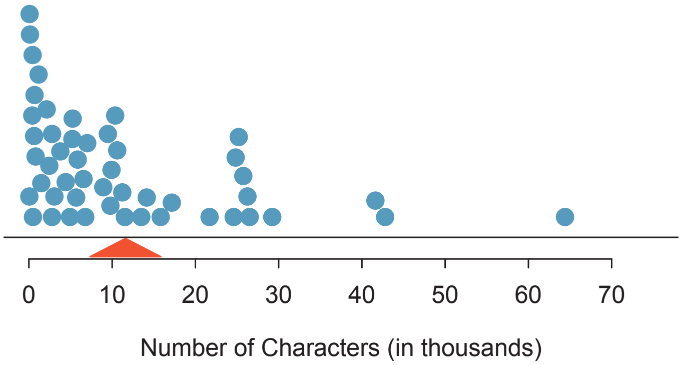

Figure 1.19: A stacked dot plot of `num_char` for the `email50` data set.

Mean

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

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.19}$$

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

n
sample size

○ **Exercise 1.20** Examine Equations (1.18) and (1.19) above. What does x_1 correspond to? And x_2? Can you infer a general meaning to what x_i might represent?[24]

○ **Exercise 1.21** What was n in this sample of emails?[25]

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 1.22** 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 4 and beyond, we will develop tools to characterize the accuracy of point estimates, and we will find that point estimates based on larger samples tend to be more accurate than those based on smaller samples.

[24] 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.

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

● **Example 1.23** 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 1.23 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

1.6.3 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 email50 data set, we create a table of counts for the number of cases with character counts between 0 and 5,000, then the number of cases between 5,000 and 10,000, and so on. Observations that fall on the boundary of a bin (e.g. 5,000) are allocated to the lower bin. This tabulation is shown in Table 1.20. These binned counts are plotted as bars in Figure 1.21 into what is called a **histogram**, which resembles the stacked dot plot shown in Figure 1.19.

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 1.20: The counts for the binned num_char data.

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

Histograms are especially convenient for describing the shape of the data distribution. Figure 1.21 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**.[26]

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**.

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

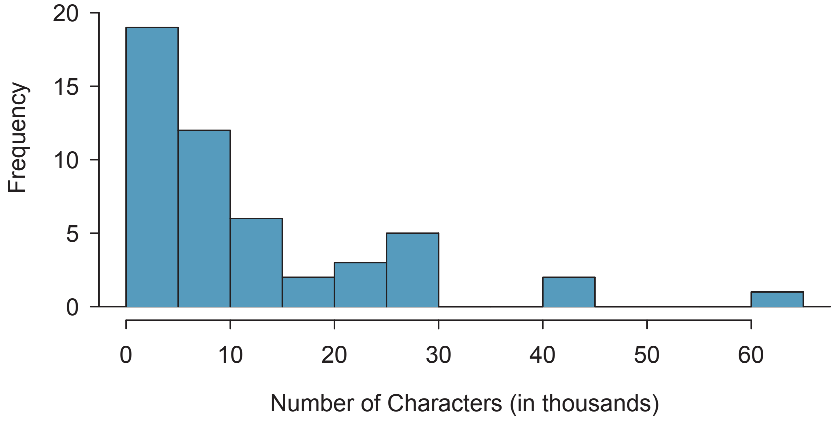

Figure 1.21: A histogram of num_char. This distribution is very strongly skewed to the right.

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

⊙ **Exercise 1.24** Take a look at the dot plots in Figures 1.18 and 1.19. Can you see the skew in the data? Is it easier to see the skew in this histogram or the dot plots?[27]

⊙ **Exercise 1.25** Besides the mean (since it was labeled), what can you see in the dot plots that you cannot see in the histogram?[28]

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.[29] There is only one prominent peak in the histogram of num_char.

Figure 1.22 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.

⊙ **Exercise 1.26** Figure 1.21 reveals only one prominent mode in the number of characters. Is the distribution unimodal, bimodal, or multimodal?[30]

[27] 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.

[28] Character counts for individual emails.

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

[30] 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.)

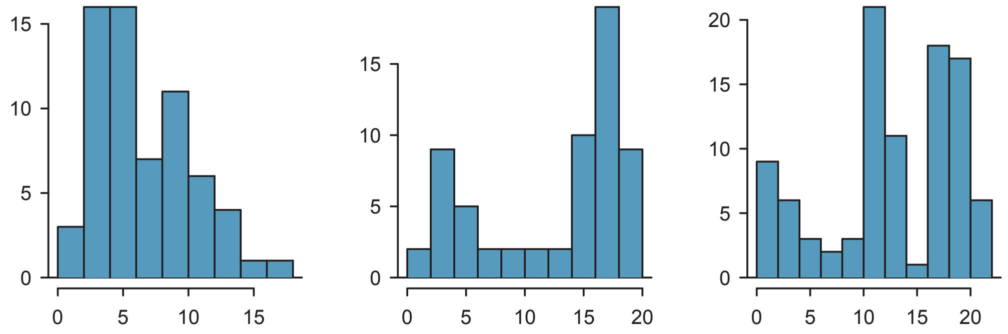

Figure 1.22: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

Exercise 1.27 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?[31]

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.

1.6.4 Variance and standard deviation

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

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

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

[31] 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.

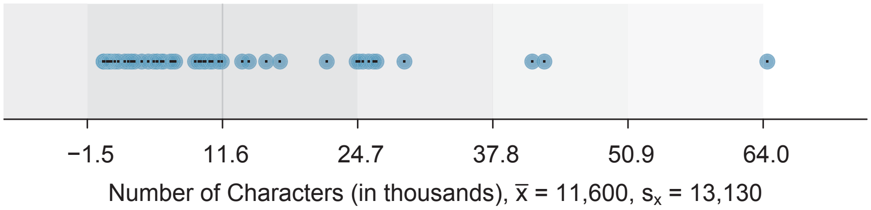

Figure 1.23: In the `num_char` data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule of thumb is less accurate for skewed data, as shown in this example.

s^2 sample variance

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

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

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

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

s 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, ..., x_n. The $_x$ subscript is usually omitted when it is clear which data the variance or standard deviation is referencing.

Variance and standard deviation

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

Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.[32] 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

σ population standard deviation

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

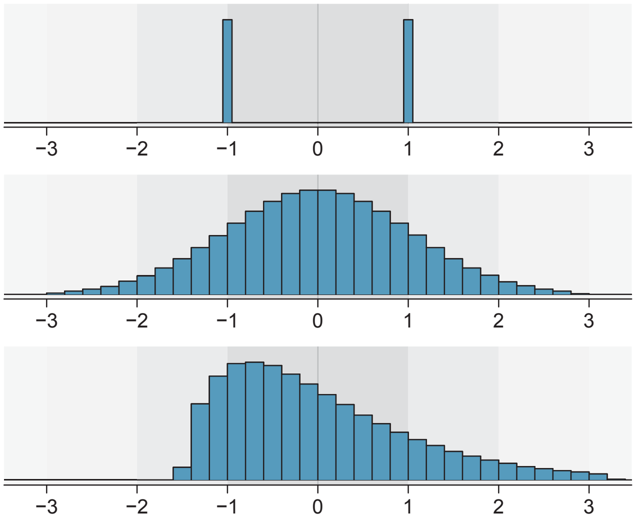

Figure 1.24: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

TIP: standard deviation describes variability
Focus on the conceptual meaning of the standard deviation as a descriptor of variability rather than the formulas. Usually 70% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, as seen in Figures 1.23 and 1.24, these percentages are not strict rules.

⊙ **Exercise 1.28** On page 23, 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 1.24 as an example, explain why such a description is important.[33]

● **Example 1.29** Describe the distribution of the `num_char` variable using the histogram in Figure 1.21 on page 24. The description should incorporate the center, variability, and shape of the distribution, and it should also be placed in context: the number of characters in emails. Also note any especially unusual cases.

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

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

[33] Figure 1.24 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.