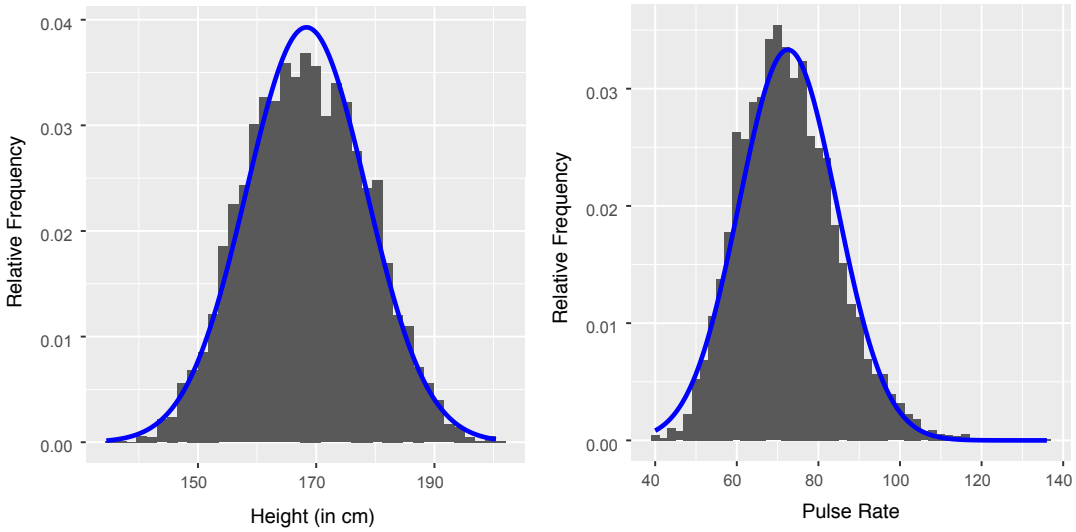

Figure 3.6: Histograms for height (left) and pulse rate (right) in the NHANES dataset, with a normal distribution curve overlaid on each dataset.

Chapter 3

Summarizing data

3.3 Idealized representations of distributions

Datasets are like snowflakes, in that every one is different, but nonetheless there are patterns that one often sees in different types of data. This allows us to use idealized representations of the data to further summarize them. Let's take the adult height data plotted in 3.5, and plot them alongside a very different variable: pulse rate (heartbeats per minute), also measured in NHANES (see Figure 3.6).

While these plots certainly don't look exactly the same, both have the general characteristic of being relatively symmetric around a rounded peak in the middle. This shape is in fact one of the commonly observed shapes of distributions when we collect data, which we call the *normal* (or *Gaussian*) distribution. This distribution is defined in terms of two values (which we

call *parameters* of the distribution): the location of the center peak (which we call the *mean*) and the width of the distribution (which is described in terms of a parameter called the *standard deviation*). Figure 3.6 shows the appropriate normal distribution plotted on top of each of the histrograms.You can see that although the curves don't fit the data exactly, they do a pretty good job of characterizing the distribution – with just two numbers!

As we will see later in the course when we discuss the central limit theorem, there is a deep mathematical reason why many variables in the world exhibit the form of a normal distribution.

3.3.1 Skewness

The examples in Figure 3.6 followed the normal distribution fairly well, but in many cases the data will deviate in a systematic way from the normal distribution. One way in which the data can deviate is when they are asymmetric, such that one tail of the distribution is more dense than the other. We refer to this as "skewness". Skewness commonly occurs when the measurement is constrained to be non-negative, such as when we are counting things or measuring elapsed times (and thus the variable can't take on negative values).

An example of skewness can be seen in the average waiting times at the airport security lines at San Francisco International Airport, plotted in the left panel of Figure 3.7. You can see that while most wait times are less than 20 minutes, there are a number of cases where they are much longer, over 60 minutes! This is an example of a "right-skewed" distribution, where the right tail is longer than the left; these are common when looking at counts or measured times, which can't be less than zero. It's less common to see "left-skewed" distributions, but they can occur, for example when looking at fractional values that can't take a value greater than one.

3.3.2 Long-tailed distributions

Historically, statistics has focused heavily on data that are normally distributed, but there are many data types that look nothing like the normal distribution. In particular, many real-world distributions are "long-tailed",

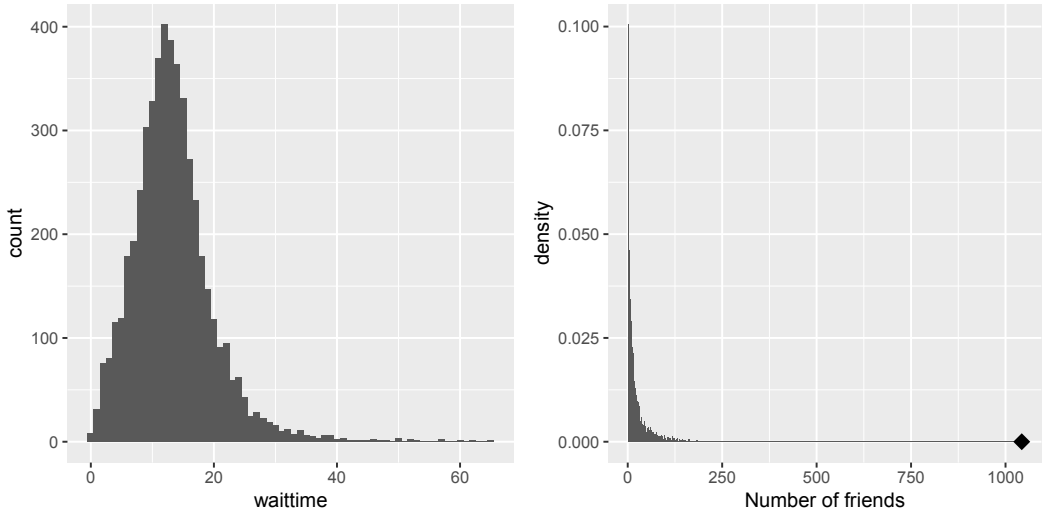

Figure 3.7: Examples of right-skewed and long-tailed distributions. Left: Average wait times for security at SFO Terminal A (Jan-Oct 2017), obtained from https://awt.cbp.gov/ . Right: A histogram of the number of Facebook friends amongst 3,663 individuals, obtained from the Stanford Large Network Database. The person with the maximum number of friends is indicated by the diamond.

meaning that the right tail extends far beyond the most typical members of the distribution. One of the most interesting types of data where long-tailed distributions occur arises from the analysis of social networks. For an example, let's look at the Facebook friend data from the Stanford Large Network Database and plot the histogram of number of friends across the 3,663 people in the database (see right panel of Figure 3.7). As we can see, this distribution has a very long right tail – the average person has 24.09 friends, while the person with the most friends (denoted by the blue dot) has 1043!

Long-tailed distributions are increasingly being recognized in the real world. In particular, many features of complex systems are characterized by these distributions, from the frequency of words in text, to the number of flights in and out of different airports, to the connectivity of brain networks. There are a number of different ways that long-tailed distributions can come about, but a common one occurs in cases of the so-called "Matthew effect" from the Christian Bible:

> For to every one who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away. — Matthew 25:29, Revised Standard Version

This is often paraphrased as "the rich get richer". In these situations, advantages compound, such that those with more friends have access to even more new friends, and those with more money have the ability to do things that increase their riches even more.

As the course progresses we will see several examples of long-tailed distributions, and we should keep in mind that many of the tools in statistics can fail when faced with long-tailed data. As Nassim Nicholas Taleb pointed out in his book "The Black Swan", such long-tailed distributions played a critical role in the 2008 financial crisis, because many of the financial models used by traders assumed that financial systems would follow the normal distribution, which they clearly did not.