# 1 — Introduction

In the 1990s and before, most of the world's information was stored on paper and other analog media, such as film. However, with the proliferation of personal computers and the internet, by 2000 one-quarter of the world's information was stored digitally. Since that time, the amount of digital data has exploded, roughly doubling every couple of years, so that now *more than 98% of all stored information is digital*.

Much of this digital data is the result of the *datafication* of the world. Datafication is both the digitization of existing analog media and, more significantly, the collection of digital data on people, processes, and other things in ways that until recently were not possible. For example, the rise of social media has resulted in the generation of massive amounts of digital data by and about individuals throughout the world. More generally, the proliferation of smart sensors and ever-cheaper storage is driving the availability of data from all types of societal, commercial, and government processes and systems. The result is an exponentially increasing amount of data being collected and stored, much of which is in need of analysis so that useful information can be extracted from the data.

What does some of this data look like? In May 2018, Bernard Marr, writing for *Forbes* magazine, said

> The amount of data we produce every day is truly mind-boggling. There are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating with the growth of the Internet of Things (IoT). Over the last two years alone 90 percent of the data in the world was generated. This is worth re-reading!

The information cited by Marr, which is now several years old, includes over 527,000 photos shared on Snapchat, 120 new professionals on LinkedIn, over 4 million YouTube videos, and over 450,000 tweets; all of these statistics are per minute! These numbers have increased since 2018 and are continuing to increase. We use Google to conduct 40,000 searches per second. One-fifth of the world's population, 1.5 billion people, use Facebook daily.

This flood of digital data contains valuable information that can be used to inform all types of decisions. Indeed, cutting-edge commercial organizations now focus their operations around the analysis and exploitation of knowledge gleaned from data. This is what data science is all about: turning data into useful information.

## 1.1 Data Science and Statistics

Both data science and statistics are concerned with the extraction of useful information from data. Let's start by defining the two fields.

**Definition 1.1.1 — Statistics.** Statistics is the science of learning from data, including collecting, organizing, analyzing, interpreting, and presenting data, often with a particular focus on measuring, controlling, and communicating the uncertainty inherent in the data, associated analyses, and final results or conclusions.

Statistics traces its roots back to 2 CE, when the Han dynasty conducted a census of the Chinese population, where it counted 57.7 million people in 12.4 million households. As this early application illustrates, statistics is about both the collection of data and its analysis. Furthermore, as Definition 1.1.1 makes clear, statistics is also focused on determining the uncertainty in data and in the conclusions drawn from data. This is an important consideration when only a sample of data is observed because the results will be subject to sampling error, and classical statistics is generally predicated on the idea that it's not possible to observe all the data, either because it's too expensive or it's impossible to collect it all. We'll explore this concept more in Section 1.2, but briefly the idea is to use the data not only to summarize what was observed but also to quantify what can be said about *all* of the data, both observed *and* unobserved.

Statistics can also be split into two broad subfields: theoretical statistics and applied statistics. Theoretical statistics is concerned with the creation and development of methods and techniques to summarize and analyze data, including clearly defining how and when to use the methods and their associated pros and cons. Applied statistics, on the other hand, is the application of the methods to data and the process of conducting rigorous and principled analyses. In the statistical profession, the division between applied and theoretical statisticians is fuzzy, with most statisticians doing both, though perhaps with an emphasis on one or the other.

The theory and practice of statistics go hand-in-hand. A rigorous theoretical foundation is what keeps statistics from being a purely empirical exercise of sifting through data and a rich set of applied problems is what keeps statistics relevant to solving real-world problems. As David Bartholomew (1995), former president of the Royal Statistical Society, said, "There can be no statistics without data and no statistics with data alone."

When developing methods, statisticians seek to understand how the methods perform on particular types of data, including how efficiently they extract information from the data and how well they characterize the uncertainty inherent in a sample of data. That is, just as automobile designers seek to understand the performance characteristics of a new car, statisticians seek to understand how their methods perform. In particular, as Lindsay et al. (2004) say, "A distinguishing feature of the statistics profession, and the methodology it develops, is the focus on a set of cautious principles for drawing scientific conclusions from data."

**Definition 1.1.2 — Data Science.** Data science is the study of how to extract useful information from data using quantitative methods and theories from many fields, including statistics, operations research, computer science, and various engineering disciplines. Data science often focuses on large data sets not originally designed or collected to address the question of interest.

In many ways data science is a modern extension of statistics and, to the extent they use statistical methods, data scientists can be characterized as applied statisticians. Indeed, while some trace the inception of data science to the 1960s, where it was then focused on data processing, modern data science originated with a lecture given by Professor C.F. Jeff Wu in 1998 entitled "Statistics = Data Science?" In that lecture, Professor Wu characterized statistical work as data modeling, analysis, and

decision making, and he proposed that statistics be renamed data science and statisticians be called data scientists.

However, since the late 1990s, particularly with the explosion of massive, heterogeneous, and often unstructured data sets, the term data science has expanded to include the ability to collect, manage, and analyze such data. Today, data scientists are expected to be adept in both statistics and computer science, particularly as applied to extracting and manipulating large data sets, as well as to have a solid working knowledge of the field in which they are trying to answer questions. In particular, data scientists must be able to:

- find, manage, and interpret large and complex data sets;
- analyze the data, including building mathematical models; and
- present and communicate results.

An important distinction between data science and statistics is whether we wish to explain or predict.[1] The goal of a study may be to make predictions. For example, the goal might be to predict when the actual number of people with influenza is high, and the predictors might include internet searches. Here the goal is to make accurate *predictions*, not to explain what variables affect the response and the extent to which they have an effect. Other times, the goal might be to explain why something happens. For example, when a cluster of cancer cases is discovered, epidemiologists will use data to search for a cause. This is an example of using data to *explain*. Often data scientists make predictions and use methodology designed for this purpose; statisticians develop models that explain how the world works. This overgeneralizes, since both perform both tasks, but this is a useful way of looking at the differences between data science and statistics.

As a result, as Definition 1.1.2 makes clear, today's data scientists come from a variety of fields and academic backgrounds and they collect and analyze data using a variety of methods. Thus, data science now extends beyond the realm of traditional statistics that was generally focused on collecting and analyzing smaller and typically very structured number-based data sets. Yet, coming full circle, those who collected and collated the census data back in 2 CE for the Han dynasty – where collecting information on 57.7 million people was undoubtedly a huge undertaking that resulted in a massive amount of data for that era – could have been called data scientists!

## 1.2 More on Statistics

While most of the colloquial and popular media references to statistics concern the collection and summarization of a set of numbers (e.g., baseball statistics or stock market returns), real statistics is about much more than that. If the field of statistics were only concerned with describing data, that is, *descriptive statistics*, this book would conclude with Chapter 2.

Statistics is most fundamentally about methods for describing uncertainty. For example, uncertainty may arise if a data science question is about a particular *population* but data are only available on a subset of the population – a *sample*. Hence, there is uncertainty about how closely the results from the sample correspond to the results for the population. Similarly, uncertainty may arise if the data science question involves forecasting the future which, of course, can only be answered using data from the past and present.

For example, what if we wanted to know the average starting salary for a person obtaining a master's degree in data science in the United States? One way to find out would entail getting the salary information for every new data scientist in the United States and then calculating the average. The left side of Figure 1.1 illustrates this idea. However, obtaining the starting salary data for every single data scientist in the United States is probably impossible. Alternatively, we could collect the starting salary information from a *sample* of new data scientists with a master's degree and use it to *estimate* the average salary of the entire population. The right side of Figure 1.1 illustrates this idea, where the goal is to use

[1] This distinction is the topic of a paper by Shmueli (2010).

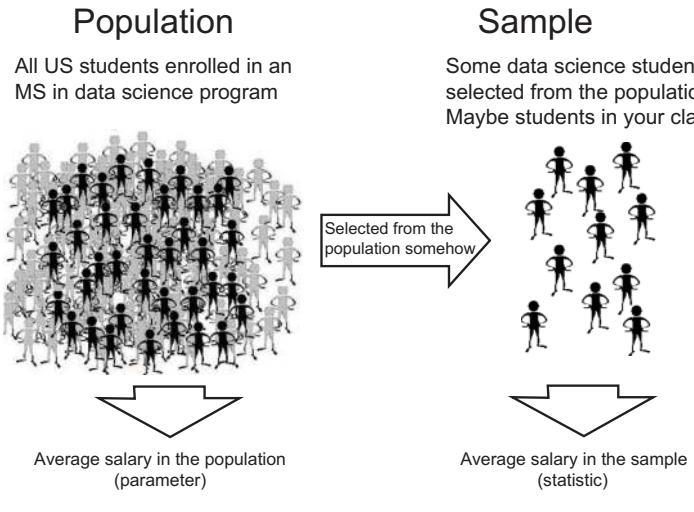


**Figure 1.1** Calculating the average starting salary for the entire population of data scientists in the United States versus the average starting salary for a sample of data scientists. The average calculated from a sample is unlikely to be the same as the average calculated from the population.

the sample results to understand the population, and so it is clearly important to ensure the sample is representative of the population.

In either situation the natural question that arises is "How far off is the estimated or predicted average salary from the actual value?" After all, in Figure 1.1 the sample is *not* the population, so any analysis on sample data is likely to differ from the same analysis done on the complete population's data. Statistical methodology is designed to formally specify the precision and uncertainty inherent in any such estimate or prediction.

Of course, since it is often difficult or impossible to know the true result for the population, it can be easy for unprincipled analysts to fool those not skilled in statistics. *Good statistics is about defining mathematically rigorous ways to do estimation, hypothesis testing, and modeling combined with principled methods for quantifying how far off an estimate is likely to be from the true answer.* That may sound like a bit of magic, but you will learn how to do it in this book!

### 1.2.1 Populations and Samples

We used the terms *population* and *sample* to motivate what statistics is all about: quantifying uncertainty. A population is the set of all people or things that meet the criteria of a particular research study or data science question. A sample is a subset of the population upon which the study or analysis will actually be done. A *random sample* is a subset that is not drawn in any systematic way from the population. (We'll learn more about sampling in Chapter 7.)

For example, if we were interested in saying something about the average GRE scores for graduate students studying data science this year, then the population would be all students enrolled in a data science degree program this year. A sample of that population could be the students in your statistics class. That sample is not likely to be random, however, since it systematically excludes certain groups of students (such as students enrolled in data science programs at other schools).

If we are interested in the average height of students in your statistics class, the class is the population. A sample might be all the women in the class. Is that a random sample? If we used the average of the heights of the women in the class to estimate the average height of all students in the class, would we be making a good estimate?

Why sample? Often it is either impossible or financially prohibitive to observe an entire population. In fact, sometimes even with significant resources and extraordinary effort it is difficult to accurately measure an entire population. A good example is the US Census. Every decade the US government

spends millions of dollars and puts forth significant effort trying to count every individual in the country. And, every decade, the Census is challenged for failing to accurately count certain segments of society.

Several governmental agencies collect large amounts of data in complex surveys. For example, the Centers for Disease Control and Prevention (CDC), part of the US Department of Health and Human Services, selects approximately 10,000 participants and provides thorough health data through a physical exam. The sampling design involves both stratification and cluster sampling, and analysis of the National Health and Nutrition Examination Survey (NHANES) data must account for this. The CDC also conducts the Behavioral Risk Factor Surveillance System, which interviews 400,000 adults regarding their health and risk behavior. The US Department of Justice administers the National Crime Victimization Survey twice per year. Each time about 50,000 households are selected for interview. All of these surveys involve a complicated sampling design which must be taken into account when doing an analysis. Interested readers are referred to Lohr (2010).

As it turns out, with good statistical practice we can often get as precise answers from a sample as we can from an attempt to collect data from the whole population. There are times when taking a sample can be *more* precise than trying to get the whole population. How can this be? Well, for the same amount of effort or cost one can either get precise data from the sample or imprecise data from the population. The idea is that under certain conditions it is preferable to allow for a moderate increase in *sampling error* in order to achieve a greater reduction in *measurement error.*

### 1.2.2 Descriptive versus Inferential Statistics

Descriptive statistics and data visualization are ways to numerically and graphically summarize data, whether the data are from a sample or a population. Why is this important? Think about the US Census, with its information on more than 300 million people. If we wanted to understand the economic status of people in the United States we would certainly not want to do so by looking at each and every Census record. Rather, we would use ways to describe the data in a more concise way, either through summary statistics or graphical plots of the data. That is, we would use descriptive statistics to summarize the data.

Most of the rest of this book is about *inferential statistics*, though we will have to spend quite a few chapters developing the probability tools we need to do statistical inference first. This is the machinery designed for using a *statistic* calculated from a sample of data to say something about the population. As illustrated in Figure 1.2, if it is impossible to obtain the starting salary for every data scientist in the United States, then we will have to use information from a sample to *infer* what it is for the population. However, inference is also more than using a sample average as an estimate for a population average; it

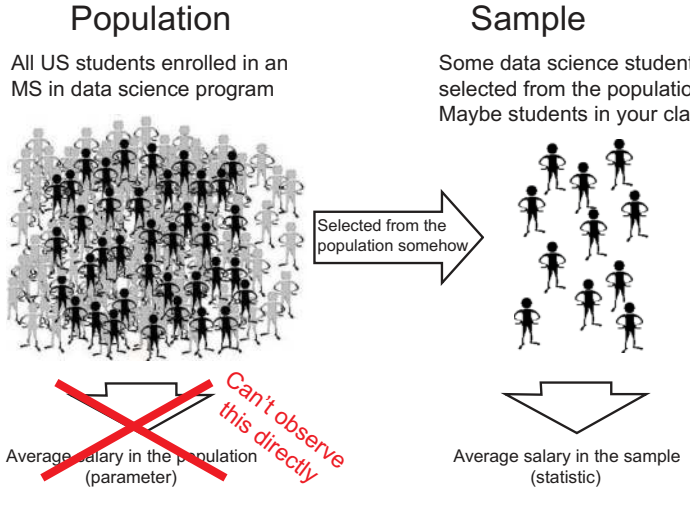


**Figure 1.2** Statistical inferential methods are designed to estimate population *parameters* using *statistics* generated from a sample and to quantify the precision of the estimate.

is also about statistical methods to quantify how accurate the sample average is and thereby specify our uncertainty in our knowledge of the population.

## 1.3 An Introduction to R

R is cutting-edge, free, open-source statistical software. R runs on a wide variety of UNIX platforms, and on the Windows and MacOS operating systems. To download it, go to `www.r-project.org` and follow the downloading and installation instructions. There are many good tutorials and books on R coding. We assume that the reader knows the basics of R and how it runs. In this book we will explain how R is used to analyze data while assuming some familiarity with R.

A common file format for data is csv, which stands for comma separated variables. Given that your data are in csv format, they can be read using the `read.csv()` function, where the argument is the file name (including the appropriate path) enclosed in quotes. It is customary to set R's working directory through the `setwd` command; for example:

```
setwd("/Users/rdfricker/Desktop")
```

Once the directory is set, R looks for files in this directory. If you are a PC user, note the use of forward slashes rather than backslashes. Also note that the default option for the `read.csv()` function is that the first line of the file contains the variable names and each subsequent line contains the data, one line for each observation (row) in the data, where each item in the data is separated from the next by a comma. If your file begins with the data in row 1 – that is, there is no row for the variable names – you can set the header to be `FALSE`:

```
my.data <- read.csv("/Users/rdfricker/Desktop/data.csv",header=FALSE)
```

You can then assign names to the columns using the `names()` function:

```
names(my.data) = c("var1","var2","var3")
```

This will give names to the first three variables in your data frame.

### Functions

One of the strengths of R is the ability to use and write functions. Functions are basically mini-programs, where you can create a function that even calls other functions. Just about everything you do in R involves applying a function, usually to data. And, as we'll discuss shortly, R users can write their own functions that can be published to the wider R community via packages that everyone can then download and use.

R has thousands of functions. We'll use many of them as we go through the text. For now, below is a list of those you need to know to get started.

- `c()` is the concatenate function, which is often used to create a vector, as in `vector1 <- c(1,2,3)` or to join two or more vectors, as in `new.vector <- c(vector1,vector1)`.
- `dim(`*`data`*`)` returns the number of rows and columns respectively in *data* which can be either a data frame object or matrix object.
- `help(`*`function.name`*`)`, `?`*`function.name`*, and `??`*`text`* are useful for getting help.
- `is.na()` is a function that returns a logical object indicating whether data are missing (i.e., NA) or not.
- `length(`*`vector`*`)` returns the number of elements in *vector*.
- `library(`*`name`*`)` loads the package called *name* so you can use/access its contents.
- `ls()` lists the objects in the workspace. No arguments are required to run the function, but you do have to type the parentheses because it is a function.

- `read.csv()` and `read.table()` are useful for reading data into R. For reading data in specialized formats, the `foreign` package is very useful.
- `rm(name)` deletes the object *name* from the workspace. It works on single objects or a series of objects separated by commas.
- `rm(list = ls(all = TRUE))` deletes all the files in your workspace. Be very cautious when using this – there is no undo option!
- `sin`, `cos`, `exp`, and `log` are, respectively, the sin, cos, exponential (i.e., $e^x$), and natural log functions. Most other mathematical functions you've encountered are built into R.

We can define and write our own functions in R using the `function` command. The following takes one argument $x$ and returns $(x+1)^2$:

```
f = function(x)
{
        y = x + 1
        z = y^2
        return( z )
}
```

This code defines the function `f` and says that it takes one argument `x`. It then computes `y = x+1` and finally squares it to obtain the variable `z`. The statement `return(z)` returns the computed value of `z` back to R. When we run the above code, nothing seems to happen, although after having run it we notice that the function is now in R's global environment; this can be seen in the upper right panel in R Studio. Once the function is executed, we can call it using the syntax `f(x)`, where `x` is some number or variable. For example, typing `f(9)` yields

```
f(9)
[1] 100
```

## 1.4 Descriptive Statistics

Descriptive statistics is all about summarizing data. In this chapter we will learn how to describe data numerically using statistics; in Chapter 2 we will learn how to graphically describe data. In fact, a *statistic* is simply a number calculated from data that summarizes something about the data. For example, the average is a statistic, and there are many other types that we will learn about in this chapter.

Descriptive statistics are becoming increasingly important as data scientists deal with ever-larger data sets and as data collection accelerates in our ever more computerized and interconnected world. They are important because the human mind is limited in its ability to assimilate individual facts; we simply aren't good at being able to synthesize lots of numbers. Indeed, human short-term memory capacity is only about seven digits – the length of a US telephone number *not* counting the area code. So, in our increasingly data-rich world, knowing how to appropriately summarize data is a critical skill.

Returning to our discussion of the US Census in Section 1.2.1, the only way to get some understanding of the US population using the Census is to apply descriptive statistics (as well as other methods we will talk about in later chapters) to summarize the data. Just looking at individual Census records would not provide us with much insight into the entire US population. And, while the 300 million plus Census records may sound like a lot, these days it is not a big data set: There are now more tweets sent *per day* than there are Census records.

Before we proceed further, let's formalize what we've just discussed with some definitions.

**Definition 1.4.1 — Data.** Information, often numerical but not necessarily so, collected from an experiment, a survey, administrative records, the internet, etc. The word "data" is plural. One piece of information is a "datum."

**Definition 1.4.2 — Statistic.** A numerical fact, usually computed from a data set. Statistics can also be computed from subsets of the data and can even be just a single datum.

**Definition 1.4.3 — Descriptive Statistic.** A statistic that usefully summarizes a data set, where the data can be either for an entire population or for a subset of the population.

Good descriptive statistics can help data scientists understand what the data are trying to say. They can highlight and bring out the underlying information in a data set, which might not (and probably will not) be evident by just inspecting the individual data elements.

### 1.4.1 Types of Data

We can divide data into two basic types, quantitative and qualitative. *Quantitative* data are data that can be measured or characterized with a numerical value; *qualitative* data cannot be so measured. For example, if we think about demographic data, height, weight, and age are all quantitative, while gender and eye color are qualitative.

We can also divide data into *cross-sectional* data and *longitudinal* (also known as *time series*) data. Cross-sectional data are data that occur either in one time period or are constant over time, while longitudinal data covers multiple time periods or varies over time. For example, referring back to the previous demographic data, gender and eye color are cross-sectional in the sense that they are unlikely to change over time. Height and weight data could be cross-sectional if they are recorded for one period of time and longitudinal if they are repeatedly measured over multiple time periods.

As shown in Figure 1.3, we can further describe quantitative data as either *continuous* or *discrete*. Data are discrete if there are gaps between the values the data can assume. For example, the number of people in a family can be 1, 2, 3, …. It cannot be 2.7. If there are no gaps between possible data values, then we say the data are continuous. Another way to think about continuous data is as if we had an infinitely accurate measuring device, then we could express the data to any number of decimal places and it would always make sense. For example, height is continuous: we can talk about someone being 6 feet tall, or 5.97 feet tall, or 5.9722683 feet tall.[2]

Unlike quantitative data, qualitative data cannot be measured or described numerically. As shown in Figure 1.3, qualitative data can be either *nominal* or *ordinal*. Ordinal qualitative data are data for which there is a natural ordering, but the data cannot be expressed on a numerical scale. For example, shirt size is ordinal: "large" shirts are bigger than "medium" shirts which are bigger than "small" shirts. In contrast, with nominal data there is no natural ordering to the data. For example, gender is a nominal type of data: each person can be classified as "male" or "female," but it does not make any sense to say that "male" is greater than "female."

Note that nominal data can be represented numerically for purposes of analysis, but care must be taken not to over-interpret the numerical labels. For example, we will sometimes use an *indicator*

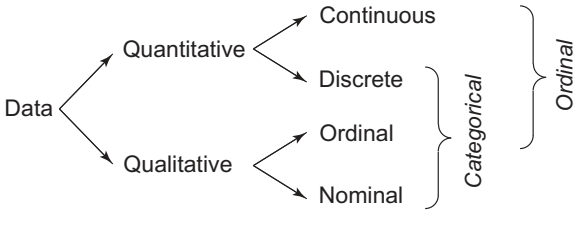


**Figure 1.3** A taxonomy of types of data.

[2] In practice, height can be measured to a fixed degree of precision, probably about one-tenth of an inch. Thus, all data are discrete because we can measure to a fixed degree of precision. Despite this, it is often helpful to think of data such as height as continuous. All models, such as how we measure height, are *approximations* of reality.

*variable* to do analyses, say setting a variable called gender equal to 1 for men and 0 for women. But just because the numbers 0 and 1 are ordinal, this property does not carry over to the original qualitative variable, and so care must be taken not to over-interpret or misuse the indicator variable.

We use the term *categorical* to refer to data that are either discrete or qualitative because we can naturally categorize these types of data into groups. Continuous data are clearly ordinal since numeric data have an obvious ordering. Finally, note that we can turn continuous data into categorical data by defining ranges of values for each category. For example, for height, people may be categorized as "short" if they are less than 5 feet tall; "average" if they are between 5 and 6 feet; and "tall" if they are 6 feet or greater. The reason these distinctions are important is that the appropriate statistical analyses, and even the proper way to display data, will depend on the type of data.

### 1.4.2 Example Data: US Domestic Flights from 1987 to 2008

To help make the ideas and methods of this chapter concrete, we will illustrate them using a data set consisting of US domestic airline flight arrivals and departures and associated details for all commercial flights from October 1987 to April 2008. This is a large data set: there are nearly 120 million records for 3,376 airports. In this chapter we'll mainly focus on the data from years 1988, 1997, and 2007, where there are 5,202,096, 5,411,843, and 7,453,215 observations (i.e., flights), respectively, for those years.

See the book's website for instructions on how to download this and other data sets. As shown in Table 1.1, the data set contains information on the date, day, and time of each flight, the airline and particular airplane, flight origin and destination, and lots of information about the length of each flight and the types of delays, if any, experienced. Table 1.2 shows five randomly selected observations from the 2007 data.

Looking through Table 1.2 we see, for example, that the time an airplane is in the air (see the `AirTime` variable) and the distance flown (see the `Distance` variable) are continuous data, while the day of the month (see the `DayofMonth` variable) is an example of discrete data. The reason for flight cancellation (see the `CancellationCode` variable) and airport of origination (see the `Origin` variable) are examples of nominal data.

Some of the data that are numerically coded are actually qualitative. For example, as we see in Table 1.2, the day of the week (see the `DayOfWeek` variable) contains the integers 1 to 7, where the number 1 corresponds to Monday, 2 to Tuesday, etc. Though these data are numbers, they are not quantitative data. Instead, the numbers are only codes that represent ordinal qualitative data. Similarly, flight numbers (`FlightNum`) are also coded numerically, but they are really nominal.

## 1.5 Cross-Sectional Data

As we just discussed, cross-sectional data are collected during the same period of time. Statistics can then be used to summarize these data.

### 1.5.1 Measures of Location

*Measures of location*, also referred to as *measures of central tendency*, are typically used to quantify where the "center" or mass of the data is located. There are a number of common measures of central tendency, each of which quantifies the "center" in a different way. The most common measure is the *mean*, which is the average of a set of observations in either a sample or a population.

**Definition 1.5.1 — Population Mean.** For data from a population, denoted $x_1, \ldots, x_N$, the population mean is calculated as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

**Table 1.1** Aircraft data set variables and brief descriptions.

| | Variable name | Description |
|---|---|---|
| 1 | `Year` | 1987–2008 |
| 2 | `Month` | 1–12 |
| 3 | `DayofMonth` | 1–31 |
| 4 | `DayOfWeek` | 1 (Monday) – 7 (Sunday) |
| 5 | `DepTime` | Actual departure time (local, hhmm) |
| 6 | `CRSDepTime` | Scheduled departure time (local, hhmm) |
| 7 | `ArrTime` | Actual arrival time (local, hhmm) |
| 8 | `CRSArrTime` | Scheduled arrival time (local, hhmm) |
| 9 | `UniqueCarrier` | Unique carrier code |
| 10 | `FlightNum` | Flight number |
| 11 | `TailNum` | Plane tail number |
| 12 | `ActualElapsedTime` | In minutes |
| 13 | `CRSElapsedTime` | In minutes |
| 14 | `AirTime` | In minutes |
| 15 | `ArrDelay` | Arrival delay, in minutes |
| 16 | `DepDelay` | Departure delay, in minutes |
| 17 | `Origin` | Origin IATA airport code |
| 18 | `Dest` | Destination IATA airport code |
| 19 | `Distance` | In miles |
| 20 | `TaxiIn` | Taxi in time, in minutes |
| 21 | `TaxiOut` | Taxi out time, in minutes |
| 22 | `Cancelled` | 1 = yes, 0 = no |
| 23 | `CancellationCode` | Reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 24 | `Diverted` | 1 = yes, 0 = no |
| 25 | `CarrierDelay` | In minutes |
| 26 | `WeatherDelay` | In minutes |
| 27 | `NASDelay` | In minutes |
| 28 | `SecurityDelay` | In minutes |
| 29 | `LateAircraftDelay` | In minutes |

**Definition 1.5.2 — Sample Mean.** For a sample of data, $x_1, \ldots, x_n$, the sample mean $\bar{x}$ is calculated as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The difference between the population mean $\mu$ and the sample mean $\bar{x}$ is subtle, but important. The population mean $\mu$ is the average across all units in the population. Usually $\mu$ is unknown, but it is important to have a symbol for it because estimating it is an important problem in statistics. The sample mean $\bar{x}$ is usually known, because it is the result of an observed sample. We often use $\bar{x}$ to estimate $\mu$. More on this idea of estimation in Chapter 9.

■ **Example 1.1 — Calculating the sample mean.** Calculate the mean for the following sample of data: {2.3, 8.1, 5.5, 9.0, 7.8}.