1

Introduction

We are surrounded by data. With a tap at a computer keyboard, we have access to more than we could possibly absorb in a lifetime. But is this data the same as information? How do we get from numbers to understanding? How do we identify simplifying trends – but also find exceptions to the rule? The computers that provide access to the data also provide the tools to answer these questions. Unfortunately, owning a hammer does not enable us to build a fine house. It takes experience using the tools, knowing when they are appropriate, and also knowing their limitations.

The study of statistics provides the tools to create understanding out of raw data. Expertise comes with experience, of course. We need equal amounts of theory (in the form of statistical tools), technical skills (at the computer), and critical analysis (identifying the limitations of various methods for each setting). A lack of one of these cannot be made up by the other two.

This chapter provides a review of statistics in general, along with the mathematical and statistical prerequisites that will be used in subsequent chapters. Even more broadly, the reader will be reminded of the larger picture. It is very easy to learn many statistical methods only to lose sight of the point of it all.

1.1 What Is Statistics?

In an effort to present a lot of mathematical formulas, we sometimes lose track of the central idea of the discipline. It is important to remember the big picture when we get too close to the subject.

Let us consider a vast wall that separates our lives from the place where the information resides. It is impossible to see over or around this wall, but every now and then we have the good fortune of having some pieces of data thrown over to us. On the basis of this fragmentary sampled data, we are supposed to infer the composition of the remainder on the other side. This is the aim of *statistical inference*.

The population is usually vast and infinite, whereas the sample is just a handful of numbers.

> In statistical inference we infer properties of the population from the sample.

There is an enormous possibility for error, of course. If all of the left-handed people I know also have artistic ability, am I allowed to generalize this to a statement that all left-handed people are artistic? I may not know very many left-handed people. In this case I do not have much data to make my claim, and my statement should reflect a large possibility of error. Maybe most of my friends are also artists. In this case we say that the sampled data is *biased* because it contains more artists than would be found in a representative sample of the population.

The population in this example is the totality of all left-handed people. Maybe the population should be *all* people, if we also want to show that artistic ability is greater in left-handed people than in right-handed people. We can't possibly measure such a large group. Instead, we must resign ourselves to the observed or *empirical* data made up of the people we know. This is called a *convenience sample* because it is not really random and may not be representative.

Consider next the separate concepts of sample and population for numerically valued data. The sample *average* is a number that we use to infer the value of the population *mean*. The average of several numbers is itself a number that we obtain. The population mean, however, is on the other side of the imaginary wall and is not observable. In fact, the population mean is almost an unknowable quantity that could not be observed even after a lifetime of study. Fortunately, statistical inference allows us to make statements about the population mean on the basis of the sample average. Sometimes we forget that this inference is taking place and will confuse the sample statistic with the population attribute.

> Statistics are functions of the sampled data. Parameters are properties of the population.

Often the sampled data comes at great expense and through personal hardship, as in the case of clinical trials of new therapies for life-threatening diseases. In a clinical trial involving cancer, for example, costs are typically many thousands of dollars per patient enrolled. Innovative therapies can easily cost ten times that amount. Sometimes the most important data consists of a single number, such as how long the patient lived, recorded only after the patient loses the fight with his or her disease.

Sometimes we attempt to collect all of the data, as in the case of a *census*. The U.S. Constitution specifically mandates that a complete census of the population be performed every ten years.[1] The writers of the Constitution knew that in order to

[1] Article 1, Section 2 reads, in part: "Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be

have a representative democracy and a fair tax system, we also need to know where the people live and work. The composition of the House of Representatives is based on the decennial census. Locally, communities need to know about population shifts to plan for schools and roads. Despite the importance of the census data, there continues to be controversy on how to identify and count certain segments of the population, including the homeless, prison inmates, migrant workers, college students, and foreign persons living in the country without appropriate documentation.

Statistical inference is the process of generalizing from a sample of data to the larger population. The sample average is a simple statistic that immediately comes to mind. The Student t-test is the principal method used to make inferences about the population mean on the basis of the sample average. We review this method in Section 2.5. The sample *median* is the value at which half of the sample is above and half is below. The median is discussed in Chapter 7.

> The standard deviation measures how far individual observations deviate from their average.

The sample *standard deviation* allows us to estimate the scale of variability in the population. On the basis of the normal distribution (Section 2.3), we usually expect about 68% of the population to appear within one standard deviation (above or below) of the mean. Similarly, about 95% of the population should occur within two standard deviations of the population mean.

> The standard error measures the sampling variability of the mean.

A commonly used measure related to the standard deviation is the *standard error*, also called the *standard error of the mean* and often abbreviated SEM. These two similar-sounding quantities refer to very different measures. The standard error estimates the standard deviation associated with the sample average. As the sample size increases, the standard deviation (which refers to individuals in the population) should not appreciably change. On the other hand, a large sample size is associated with a precise estimate of the population mean as a consequence of a small standard error. This relationship provides the incentive for larger sample sizes, allowing us to estimate the population mean more accurately. The relationship is

$$\text{Standard error} = \frac{\text{Standard deviation}}{\sqrt{\text{Sample size}}}$$

determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons. The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct."

Consider a simple example. We want to measure the heights of a group of people. There will always be tall people, and there will always be short people, so changing the sample size does not appreciably alter the standard deviation of the data. Individual variations will always be observed. If we were interested in estimating the average height, then the standard error will decrease with an increase in the sample size (at a rate of $1/\sqrt{\text{sample size}}$), motivating the use of ever-larger samples. The average will be measured with greater precision, and this precision is described in terms of the standard error. Similarly, if we want to measure the average with twice the precision, then we will need a sample size four times larger.

Another commonly used term associated with the standard deviation is *variance*. The relationship between the variance and the standard deviation is

Variance = (Standard deviation)^2

The standard deviation and variance are obtained in SAS using `proc univariate`, for example. The formula appears often, and the reader should be familiar with it, even though its value will be calculated using a computer.

Given observed sample values $x_1, x_2, \ldots, x_n$, we compute the *sample variance* from

$$s^2 = \text{sample variance} = \frac{1}{n-1} \sum_i (x_i - \overline{x})^2, \tag{1.1}$$

where $\overline{x}$ is the average of the observed values.

This estimate is often denoted by the symbol s^2. Similarly, the estimated sample standard deviation s is the square root of this estimator. Intuitively, we see that (1.1) averages the squared difference between each observation and the sample average, except that the denominator is one less than the sample size. The "$n-1$" term is the degrees of freedom for this expression and is described in Sections 2.5 and 2.7.

1.2 Statistics in the News: The Weather Map

Sometimes it is possible to be overwhelmed with too much information. The business section of the newspaper is filled with stock prices, and the sports section has a wealth of scores and data on athletic endeavors. The business section frequently has several graphs and charts illustrating trends, rates, and prices. The sports pages have yet to catch up with the business section in terms of aids for the reader.

As an excellent way to summarize and display a huge amount of information, we reproduce the U.S. weather map from October 27, 2008, in Figure 1.1. There are several levels of information depicted here, all overlaid on top of one another. First we recognize the geographic-political map indicating the shorelines and state boundaries. The large map at the top provides the details of that day's weather. The large Hs indicate the locations of high barometric pressure centers. Regions with

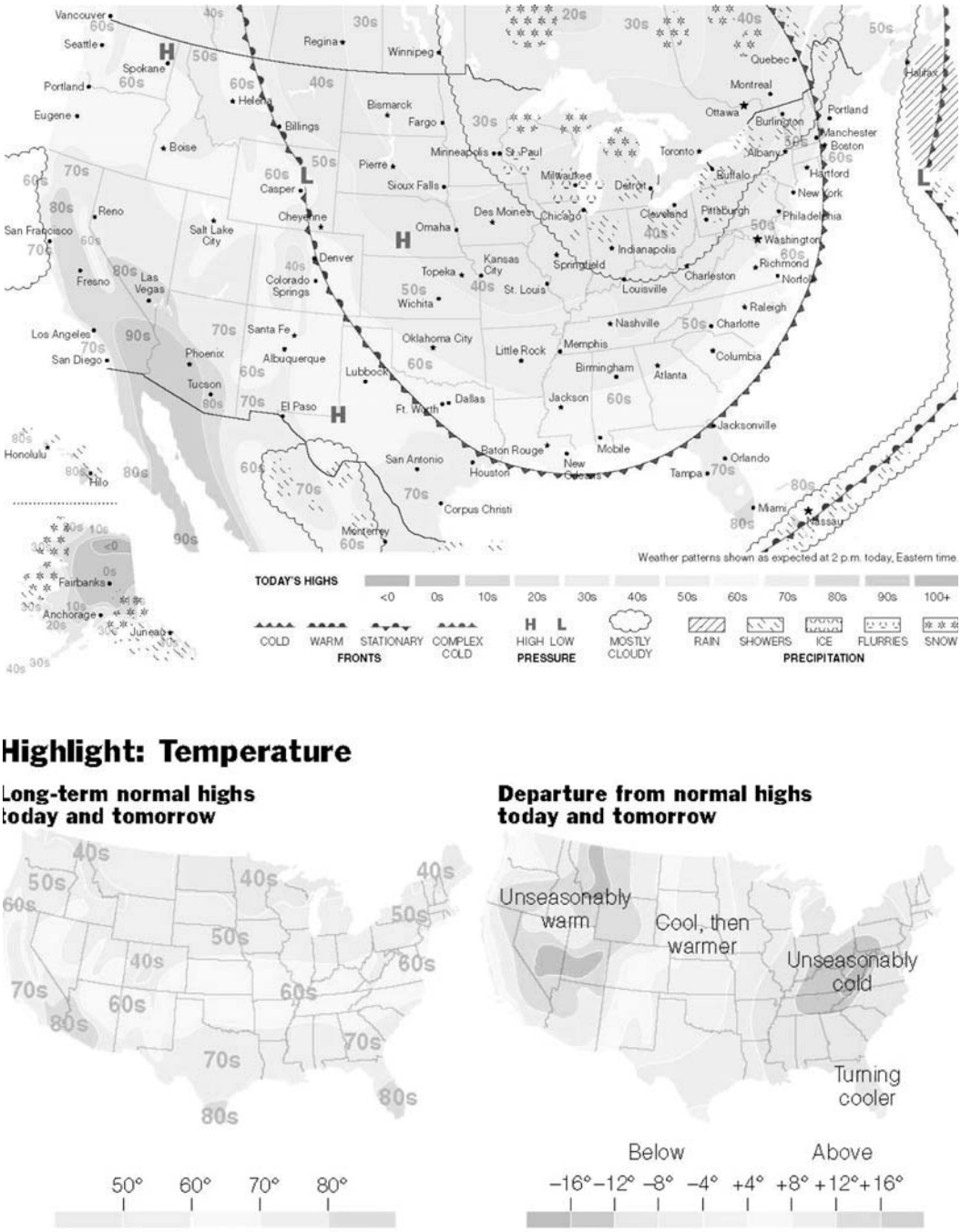

Figure 1.1 The U.S. weather map for October 27, 2008: Observed, expected, and residual data. Courtesy of Pennsylvania State University, Department of Meteorology.

similar temperatures are displayed in the same colors. The locations of rain and snow are indicated. An element of time and movement can also be inferred from this map: A large front has come across the country from the north, bringing cooler temperatures along with it. This figure represents the fine art of summarizing a huge amount of information.

The two smaller figures at the bottom provide a different kind of information. The lower-left map indicates the temperatures that we should expect to see on this date, based on previous years' experiences. The general pattern follows our preconception that southern states are warmer and northern states are cooler at this time of the year, with bands of constant temperature running east and west.

The figure on the lower right summarizes the differences between the normal pattern and the temperatures given in the large map at the top. Here we see that Florida is much cooler than what we would expect for late October. Similarly, Montana is cold at this time of year but is much warmer than typical.

The aim of statistics is to provide a similar reduction of a large amount of data into a succinct statement, generalizing, summarizing, and providing a clear message to your audience.

> The goal of statistics is to start with the data and then prepare a concise summary of it.

1.3 Mathematical Background

We all need to start someplace. Let us decide on the common beginning point.

Many of us chose to study the health or social sciences and shunned engineering or physics in order to avoid the abstract rigor of mathematics. However, much of the research in the social and health fields is quantitative. We still need to demonstrate the benefit of any proposed intervention or social observation.

For example, we all know the role that the ASPCA and other animal shelters perform in protecting homeless cats and dogs. It only takes a quick visit to their local facilities to assess the effectiveness of their efforts. We can easily count the number of charges under their care to quantify and measure what they do. In this example it is easy to separate the emotional appeal from the quantity of good such an organization supplies.

In contrast, we are shocked to see the brutality of whales being slaughtered. We are told about the majesty of their huge size and life under the sea. This is all fine and plays on our emotions. Before we send money to fund the appropriate charity, or decide to enforce global bans on whaling, we also should ask how many whales there are, and perhaps how this number has changed over the past decade. This information is much harder to get at and is outside our day-to-day experiences. We need to rely on estimates to quantify the problem. Perhaps we also need to question

who is providing these estimates and whether the estimates are biased to support a certain point of view. An objective estimate of the whale population may be difficult to obtain, yet it is crucial to quantifying the problem.

As a consequence, we need to use some level of mathematics. The computer will do most of the heavy lifting for us, but we will also need to understand what is going on behind the scenes. We need to use algebra and especially linear functions. So when we write

$$y = a + bx,$$

we recall that a is referred to as the *intercept* and b is called the *slope*. We need to recognize that this equation represents a straight-line relationship and be able to graph this relationship.

We will need to use logarithms. Logarithms, or logs for short, are always to the base $e = 2.718\ldots$ and never to base 10. The exponential function written as e^x or $\exp(x)$ is the inverse process of the logarithm. That is,

$$\log(e^x) = x$$

and

$$e^{\log x} = \exp(\log x) = x.$$

Sometimes we will use the exponential notation when the argument is not a simple expression. It is awkward to write

$$e^{a+bw+cx+dy},$$

not to mention that it is difficult to read and that publishers hate this sort of expression.

It is easier on the reader to write this last expression as

$$\exp(a + bw + cx + dy).$$

1.4 Calculus

For those who took calculus a long time ago and have not used it since, the memories may be distant, fuzzy, and perhaps unpleasant. Calculus represents a collection of important mathematical tools that will be needed from time to time in our discussion later on in this book. We will need to use several useful results that require calculus.

Fortunately, there is no need to dig out and dust off long-forgotten textbooks. The actual mechanics of calculus will be reviewed here, but there will not be a need to actually perform the mathematics involved. The reader who is fluent in the relevant mathematics may be able to fill in the details that we will gloss over.

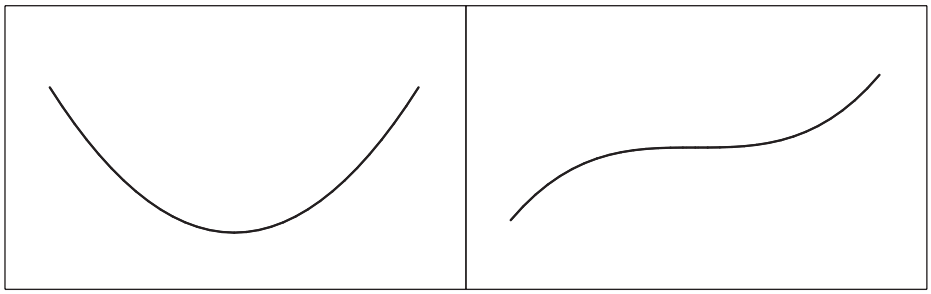

Figure 1.2 The slope is zero at the minimum of a function (left) and also at the saddle point of a function (right).

What is the point of calculus? If x and y have a straight-line relationship, we should be familiar with the concept of the *slope* of the line. When x changes by one unit, the slope is the amount of change in y.

For a nonlinear relationship, the concept of the slope remains the same, but it is a more local phenomenon. The idea of the slope depends on where in the x–y relationship your interest lies. At any point in a curve, we can still talk about the slope, but we need to talk about the slope at each point of the curve. You might think of a curve as a lot of tiny linear segments all sewn together, end to end. In this case, the concept of slope is the ratio of a small change in y to the resulting small change in x at a given point on the curve. It still makes sense to talk about the ratio of these small amounts resulting in a definition of the slope of a curved line at every point x. In calculus, the *derivative* is a measure of the (local) slope at any given point in the function.

The derivative of a function provides its slope at each point.

The derivative is useful for identifying places where nonlinear functions achieve their minimums or maximums. Intuitively, we can see that a smooth function that decreases for a while and then increases has to pass through some point where the slope is zero. Solving for the places where the derivative is zero tells us where the original function is either maximized or minimized. See Figure 1.2 for an illustration of this concept.

Some functions also exhibit *saddle points* where the derivative is also zero. A saddle point is where an increasing function flattens out before resuming its increase. We will not concern ourselves with saddle points. Similarly, a zero value of the derivative may only indicate a local minimum or maximum (that is, there are either larger maximums or smaller minimums someplace else), but we will not be concerned with these topics either. A saddle point is illustrated in Figure 1.2.

Although we will not actually obtain derivatives in this book, on occasion we will need to minimize and maximize functions. When the need arises, we will recognize

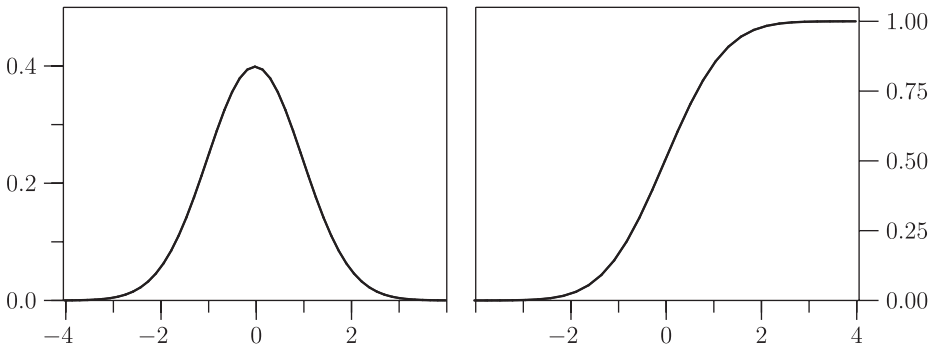

Figure 1.3 The normal density function (left) and its cumulative area (right).

the need to take a derivative and set it to zero in order to identify where the minimum occurs.

> The function achieves a maximum or minimum where the derivative is zero.

Calculus is also concerned with *integrals of functions*. Briefly, an integral gives us the area between the function and the horizontal axis. As with the derivative, we will not actually need to derive one here. Many probabilities are determined according to the area under a curved function.

> The integral of a function provides the area between the curve and the horizontal x axis.

Specifically, when we examine the normal distribution (Section 2.3), we will often draw the familiar bell-shaped curve. This curve is illustrated in Figure 1.3. For any value x on the horizontal axis, the curve on the right gives us the cumulative area under the left curve, up to x. The total area on the left is 1, and the cumulative area increases up to this value. The cumulative area under this curve is almost always of greater interest to us than the bell curve itself. Table A.1 in the appendix provides this area for us. It is very rare to see a table of the bell curve.

The area can be negative if the function is a negative number. Negative areas may seem unintuitive, but the example in the following section illustrates this concept.

1.5 Calculus in the News: New Home Sales

Home sales and building starts for new homes are both an important part of the economy. Builders will not start an expensive project unless they are reasonably sure that their investment will pay off. Home buyers will usually also purchase new furniture and carpets and will hire painters and carpenters to remodel as they move

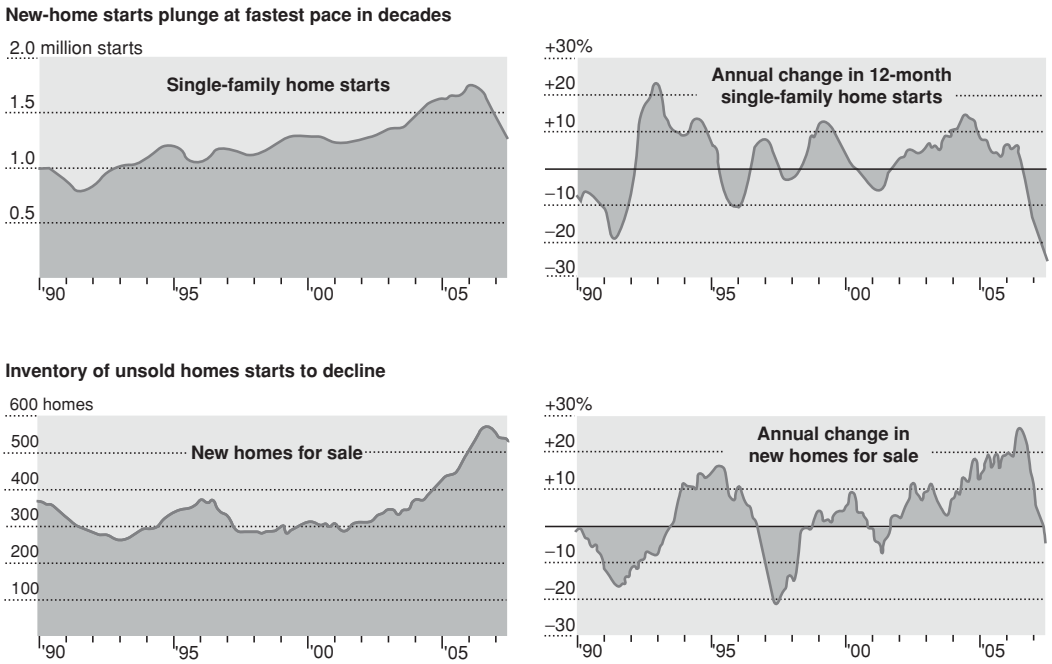

Figure 1.4 New home starts and sales. *Source: New York Times*.

in. Investors, economists, and government policy makers watch this data as a sign of the current state of the economy as well as future trends.

The graphs in Figure 1.4[2] depict new single-family home starts (upper left) and the number of new homes already on the market (lower left) over a period of a decade. There are always new homes being built and put up for sale, of course, but it is useful to know whether the trend is increasing or decreasing. The graphs on the right half of this figure show the trend more clearly in terms of the annual changes. More specifically, the graphs on the right show the slope of the line on the left at the corresponding point in time. When the figure on the left is increasing, then the figure on the right is positive. Decreasing rates on the left correspond to negative values on the right.

In words, the graphs on the right half of this figure are the derivatives of the graphs on the left half. Similarly, if we start at the values corresponding to the start of the year 1990, then the graphs on the left half are obtained by integrating the values on the right. Areas under the negative values on the right integrate to "negative areas" so that negative values on the right correspond to declining values on the left.

The times at which the derivatives on the right are zero correspond to turning points where maximums or minimums occur on the left. Remember that a zero slope is usually indicative of a change in direction. These maximums or minimums

[2] The graphs are available online at http://www.nytimes.com/2007/06/23/business/23charts.html.