CHAPTER 1: INTRODUCTION

Most research begins with a question about the world. To answer the question, *data* are collected. Data typically consist of numbers—often lots of numbers. The researcher's challenge then becomes one of using the information in the numbers to make conclusions about whatever question the research was designed to answer. This book is a primer about how to do that. As we shall see, the process of *statistics* entails two main branches. Description statistics is the process of condensing the many numbers that comprise the data into fewer numbers—or graphical representations of numbers—that can be comprehended more easily by the human brain. *Inferential statistics* is the process of using the data to make conclusions about whatever question prompted the research.

This chapter provides a brief synopsis of some of the main concepts that we will cover in the rest of the book book. In going through this example, you will see many instances of assertions like, "…and we will discuss this concept in more detail later in the book." This is true: in this introductory chapter, we are short on mathematics, formalism, and detail. Instead, we are going for the broad picture and we try to appeal to your intuition and common sense within the context of a specific (although somewhat whimsical) example. We try also to introduce, again intuitively, some of the major concepts that will form the bulk of the remainder of the book. So as you make your way through the remainder of the chapter, remember that your goal is to get a general feel for what's going on. For the moment anyway, don't worry much about the details—there'll be plenty of time for that later.

Estatia

In the mythical country of Estatia, consumer protection plays a leading role in tradition and culture. Accordingly, the Estatian Department of Consumer Protection (DCP) is a powerful and busy organization, always on the lookout to ensure that consumers are being treated fairly by the many large companies that bespeckle the land.

Suppose one day, that the DCP receives a complaint: an Estatian consumer group expresses suspicion that one of these companies, the Acme Corporation's Soup Division, is cheating its customers by shorting the soup volume in Acme's one-liter (1,000-ml) soup cans. That is, according to the complaint, the manufacturing process that produces the one-liter cans is filling them with, on the average, slightly less than one liter of soup.

How should the DCP go about testing the validity of this complaint? Although it has the capacity to measure the capacity of any given can with great precision, the DCP can't, of course, measure every can that exists. But they can address the problem in a well known fashion. They can select a *random sample* of soup cans from supermarket shelves across the country, measure the volume of soup in each of the sampled cans, and then use the measurements thus obtained from the sample to infer what's going on more generally.

Does Acme Short its Cans? Data and Conclusions from Samples

To see how this might be done and what the resulting conclusions might be, suppose that the DCP obtains a random sample of cans and measures the volume of soup in each can. Table 1.1 shows four possible outcomes, Cases I, II, III, and IV. Let's go through them and consider how much the data from each case would convince us that Acme is indeed shorting its one-liter cans.

Table 1.1. Weights (in ml) for Random Samples of Acme One-Liter Soup Cans.				
	Case I: n = 5	**Case II: n = 5**	**Case III: n = 5**	**Case IV: n = 100**
	978	878	990	978
	969	869	984	969
	988	888	991	988
	993	893	988	993
	1,012	912	989	1,012
				978
				992
				.
				.
				.
Mean:	M = 988	M = 888	M = 988	M = 988
Size of effect:	Small	Large	Small	Small
Variability:	Large	Small	Large	Large
Sample size:	Small	Small	Small	Large
Does Acme short its cans?	Can't tell	Probably	Probably	Probably
Confidence Interval: Where is μ?	988±14, i.e., 974 - 1,002	888±14, i.e., 874 - 902	988±2, i.e., 986 - 988	988 ± 3, i.e., 985 - 991

Case I

Suppose first that the data come out as depicted by Case-I data from a sample of n=5 cans (as you can see, we use "n" here and in general to denote the size of a sample).

The first thing you might notice is that these five cans don't all have the same volume; indeed the volumes of the five cans range from 969 to 1,012 ml. Why is this? Why don't all the cans have identical volumes? Well, variability is just part of the manufacturing process—like all manufacturing processes, this one produces products that, while nominally the same, are not all exactly identical to one another. As we shall see, such random variability—in soup can volumes, in scientific data, and in the world in general—is one of the primary obstacles to unambiguous conclusions from data and therefore its existence is one of the primary reasons that you're reading this book. We will return to a detailed consideration of variability in later chapters.

Meanwhile though, what might we conclude from these five Case-I numbers? As a first step we might condense the information in them into a single number that represents them all together, i.e., into an *average*. At the bottom of the Case-I column we have indicated the most common form of average, the *arithmetic mean*—the sum of the scores (4,940 in this instance) divided by the number of scores (5 in this instance)—which turns out to be 988.

So on average, the five cans in this Case-I sample contain 988 ml, or 12 ml less than the 1,000 ml that Acme advertises. Should this raise suspicion in the Estatian DCP researchers that Acme is shorting its soup cans? Well maybe, but the evidence seems slim. We note, for instance, that one of the cans has almost the correct amount—993 ml—and another can has even *more* than 1,000 ml. So intuitively, we'd conclude that, although the Case-I data are *consistent* with Acme's shorting its one-liter cans a little bit, the less-than-1,000 mean might just come about as a result of random variation in the particular cans that we've happened to choose for our sample.

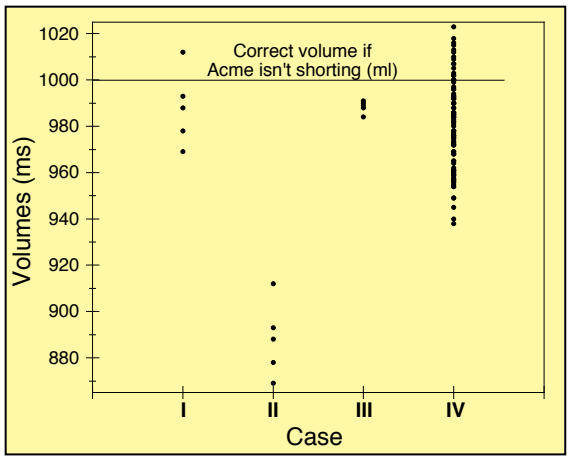

Figure 1.1. A graphical representation of the Table-1.1 data. The dots over each Case number, I-IV represent the individual data points (sample volumes in ml) from that case. The cases may be contrasted with respect to their means (Case I, III, and IV versus Case II), their variabilities (Cases I, II, and IV versus Case III) and their sample sizes (Cases I, II, and III versus Case IV).

Excel Aside. In the example, the numbers in Table 1.1 represent, of course, can volumes. But how did we create these numbers for this book? The answer is that we used a technique that is useful for acquiring insight about statistics in many situations: that of *random generation*. This technique is centered on Excel's RAND() function, which generates a random number between 0 and 1. Of course, we had to do more than that to actually generate the Table-1.1 numbers, but we will defer a description of this additional work to a later chapter.

Graphing the data

At this point, we'll take a brief side trip to introduce one of the most important axioms in the understanding of statistics, and indeed, almost anything involving numbers: *the best way of understanding numerical information is almost always to represent the numbers in some sort of pictorial form*, usually to graph them. Graphing is an art in the sense that the quality of the numerical representation is limited only by the imagination and creativity of the grapher. Whereas tables, such as Table 1.1 have their places—for instance in a textbook, one may want to provide access to a set of exact numbers—figures are almost always better than tables for conveying gut-level intuition. In this book, we will introduce numerous examples of graphing data in quest of optimal understanding.

And we do so beginning now. Figure 1.1 shows a graphical representation of the Table 1.1 data. For the moment, focus on Case I, on the left of the abscissa (horizontal axis). Above the "I" are five data points representing the five Case-I volumes from Table 1.1. We have also provided a horizontal line extending across from a volume of 1,000, which depicts the volume we'd expect assuming Acme is not shorting its cans. It will be useful to refer back to Figure 1.1 when considering the remaining three cases.

Case II

Now let's suppose that the data turned out as depicted in Case II which, as with all Cases, I-IV, are shown in both tabularly (Table 1.1) and visually (Figure 1.1.) How might the DCP's conclusions differ if the data turned out as in Case II compared to Case I? As you can see there's still variability in the Case-II data, indeed it was constructed so as to have exactly the same amount of variability as in Case I. However, the Case-II mean is 100 ml less than the Case-I mean, i.e., it's 888 rather than 988 ml. In fact, as you can see, each of the five volumes is exactly 100 ml less than its Case-I counterpart and every single sample member is below the 1,000 ml that we'd expect if Acme weren't shorting its cans. Possibly this still may not be enough to absolutely convince you that Acme is cheating, but at the very least it should be *more* convincing than the Case-I data.

Case III

Moving along, let's consider the data depicted in Case III. The mean, 998, is the same as it was in Case I. Close inspection, however, reveals an important difference: there is less variability in the five cans' weights than was true with the Case-I data. In fact, there is hardly any variability

at all: within the precision of our measuring instrument, all five cans have very close to the same volume. What should we make of this? Well, our reasoning might go: Because all five of our sample cans are very close to 988 ml, we might suppose that all of the other population cans would also have volumes close to 988 ml. And, if that were true, then the entire population would have a mean close to 988 ml. So again, although this may not constitute absolutely convincing evidence that Acme is shorting its cans, it is intuitively more so than the Case-I data which have the same mean but more variability.

Case IV

Finally, let's have a look at the Case-IV data. The big change here is that the sample is bigger—instead of sampling 5 cans, the DCP has sampled quite a few more, 100 cans. Note from Table 1.1 that the first five cans in the sample have identical volumes to the five of Case I. Looking a bit more carefully at the data, it appears that the amount of variability is about the same in Case IV as it is in Case I (and indeed it is; again the numbers were generated that way). The mean of these 100 volumes is 988 ml, just as it was in Cases I and III. Again appealing to intuition, it seems that, although again not absolutely convincing, the Case-IV data provide more evidence that Acme is shorting its cans than do the Case-I data. The reason for this is that, all else equal, we have more faith in the results of the larger, Case-IV sample than in the results of the smaller Case-I sample—if, for instance, the sample size were so big that it comprised essentially the entire population, our conclusion would be unambiguous. More generally, all else (i.e., mean and variability) being equal, the larger the sample size, the more convincing the conclusion.

A Summary and a Pseudo-Equation

What have we learned from all this? The Case-I data comprise our baseline. Given the Case-I data, we conclude that the evidence for Acme's shorting their cans, while not nonexistent, isn't really very convincing.

The data from Cases II, III, and IV all, for different reasons, provide greater evidence for shorting than do the Case-I data. To recapitulate:

Case II provides more evidence because the mean volume is lower. Let's state this in a slightly different way: The magnitude of the lowered-mean *effect* implied by shorting is greater.

Case III provides more evidence because there is less variability in the numbers.

Case IV provides more evidence because the sample size is greater.

In short, three properties of our sample—"effect" size indicated by the sample mean, variability of the sample numbers, and n, the sample size—all contribute systematically to our belief that Acme is shorting its one-liter cans. The comparative states of these three properties are indicated at the bottom of Table 1.1, and we can summarize their contributions to conclusions in the form of what we will dub a "pseudo-equation"—an equation-like formulation that comprises only rough, intuitive concepts rather than actual numbers. It is,

$$\text{Belief in shorting} = \frac{\text{Magnitude of effect}}{\left(\frac{\text{Amount of variability}}{\text{n = sample size}}\right)}$$

So the idea here is that the DCP's eventual "Belief in shorting" is influenced by the three factor we've just sketched: It's greater to the degree that "Magnitude of effect" is bigger (consider Case II compared to Case I), to the degree that "Amount of variability" is low (Case III compared to Case I) and to the degree that sample size is large (Case IV compared to Case I). In later chapters, we will formalize this pseudo-equation in various ways.

Samples, Populations, and Hypotheses

At this point it will be useful to provide some definitions and terminology. In general, we want to make conclusions about a *population*—in this instance, about the population of Acme one-liter soup cans that are sold in Estatia. To do this, we collect data from a *random sample* drawn from the population, just as we have in this example, and we use these sample data to infer what's going on with the population. There are many characteristics of a population, about which we might draw inferences from the corresponding characteristics of the sample data. In this instance, we are interested in the *population mean*. The population mean is a somewhat abstract entity that refers to the mean value of all Acme one-liter cans that exist, ever have existed, or ever could exist, using Acme's present manufacturing process. Equipped with this definition, we can reformulate the conclusion that "Acme is shorting its one-liter cans" to, "the population mean of Acme one-liter cans is less than 1,000 ml."

The term "population mean" is cumbersome. Fortunately, there's a shorter term for it: the Greek version of the letter M, or μ. Likewise, a sample mean is, as indicated in Table 1.1, simply referred to as M. Now we can formalize the DCP's decision-making task as one of deciding between two competing hypotheses:

- The "chance-effect" hypothesis: The population mean, μ, of Acme one-liter cans equals 1,000 ml. Any "effect" observed from our sample that is discrepant with this hypothesis—in this case in the form of a sample mean, M, turning out to be less than 1,000 ml—comes about just by chance.
- The "real-effect" hypothesis: The population mean, μ, of Acme one-liter cans is less than 1,000 ml. Any "effect" observed—in this case in the form of a sample mean, M, turning out to be less than 1,000 ml—comes about, at least in part for this reason.

And with this formalism, we can now recast "Belief in shorting" in our pseudo-equation as indicating support for the "real-effect" hypothesis and against the "chance-effect" hypothesis.

Hypothesis Testing and Confidence Intervals

In the preceding section, we have described the elements of what is an enormously popular procedure in many sciences. Called *hypothesis testing*, it is used in the vast majority of studies in the social sciences as well as in other sciences, particularly medicine and ecology, as a means of transiting from the data collected in some experiment to conclusions about the question that the experiment had sought to address. Despite its popularity however, a growing cadre of scientists has questioned its usefulness. The reasons underlying this doubt will be described in detail later in this book. For the moment, we want to similarly describe the elements of an alternative means of understanding the meaning of a data set, that of *confidence intervals*.

As should be apparent, hypothesis testing involves a *binary decision*; roughly speaking, a decision is made in favor of one hypothesis or the other. Note also that as we've just seen, the conclusion involves something about population means: the population mean is either concluded to be below the advertised value of 1,000 ml or it's not so concluded. A confidence interval takes a more direct approach asking, essentially: What *is* the population mean of interest? If we knew the answer to that question—e.g., in this instance, if the DCP somehow knew the population mean of Acme's one-liter cans—then the issue would be dealt with.

The sample mean, M, and the population mean, μ

As we shall demonstrate formally in a later chapter—and, as intuitively seems quite reasonable—a sample's mean, M, is the best estimate we have of μ, the population mean of the population from which the sample was drawn. So in Cases I, II, III, and IV, the best estimates of the population mean of Acme's cans are 988, 888, 988, and 988 respectively. So in each of the four cases, according to our best estimate, Acme is indeed shorting its soup cans.

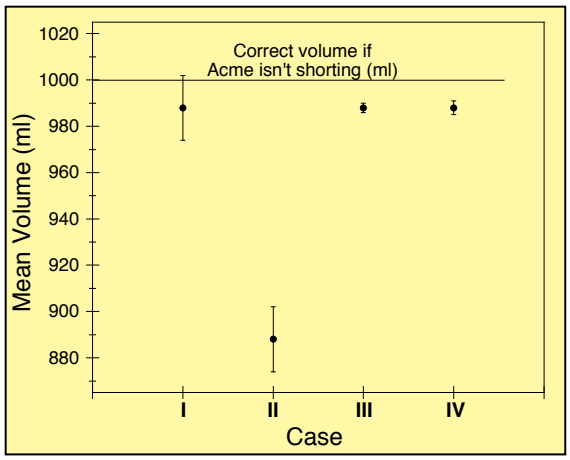

Figure 1.2. Confidence intervals. For each of Cases I-IV, we have indicated the sample mean as a black circle. The confidence interval around the mean is represented by the vertical lines extending above and below each mean. You can see that small, i.e., good confidence intervals come about with small variability (Case III) and/or a large sample size (Case IV).

Confidence intervals

How good are these estimates though? Take Case I. Although 988 ml, the sample mean is really and truly our best estimate of μ, we certainly wouldn't bet a lot of money on the proposition that μ is *exactly* 988 ml. However, we might be willing to bet on the proposition that μ is somewhere in the general vicinity of 988 ml—and in addition, the more expansive the definition of "general vicinity" the more money we'd be willing to bet.

Let's be a little more formal about these notions in the form of what's called a *confidence interval*. A confidence interval is an interval around a sample mean such that the population mean that we're concerned with falls within that interval with some pre-specified likelihood. For the moment, we'll point out a couple of things about confidence intervals.

- In case you haven't realized it already, small confidence intervals are good: the smaller the confidence interval, the more information we have about the sought-after population mean's location. In the extreme, a confidence interval of zero width, would imply that we knew perfectly where the relevant population mean lay—it would be equal to M, the sample mean.
- Confidence interval size is influenced by two of the same things that we discussed in conjunction with hypothesis testing. In a particular, a confidence interval is small to the degree that (a) the original variability is small and (b) to the degree that the sample size is large. In addition, it's influenced by the pre-specified probability that the sought-after population mean falls within the interval, i.e., a confidence interval that captures μ with 99% probability is wider than one that captures μ with 95% probability.

At the bottom of Table 1 are confidence intervals for the four cases. These same confidence intervals are also represented graphically in Figure 1.2. Take this opportunity to study these confidence intervals carefully in order to see how they correspond to the conclusions made within the context of hypothesis testing: the general idea that you should be internalizing is that those cases, II - IV, that lead us to be quite sure that Acme is shorting its soup cans are also those cases in which our confidence interval leads us to believe that the can volume population mean, μ, is less than 1,000 ml.

Probability

Across the various topics that we've been discussing runs a common thread, that of *probability*. Although we haven't stated it as such, the uncertainty with which our conclusions have been stated has implied that they are being made probabilistically rather than unequivocally. We've indicated, at least implicitly, that under this or set of circumstances, we can probably conclude that Acme is shorting its cans. Similarly, we have asserted that a confidence interval includes the sought-after population mean with some probability.

Probability theory

At first glance, this dependence on probability may seem disquieting. How are we supposed to make any kind of unambiguous conclusions if we are always hedging our bets with probabilistic statements? The answer to this question is twofold. The first answer is that because of the variability of almost everything in the world, unambiguous conclusions are almost impossible to come by. Get used to it; that's just the way things are. The second answer, a somewhat more optimistic one, is that probability is a precise mathematical discipline, which means that we can at least make probabilistic conclusions that are themselves unambiguous and that mean the same thing to everyone. To use a reasonably widely understood example, if I say that a coin is biased, such that its probability of coming up heads is 70-30 rather than the usual 50-50, that's a probabilistic statement—but it's also a meaningful one that, for instance, would be useful to you if, for instance, you were planning to toss that coin to decide who's going to pay for dinner tonight. In the next chapter, we will delve into the mathematical formalism of probability theory.

Probability and Statistics

Probability and statistics are often referred to as "two sides of the same coin." Essentially, probability is all about *deduction*. Probability comes up when we deal with the question: "Given some situation, what are the likely consequences?" So for instance if the situation tossing a fair coin six times, we might ask something like, "what is the probability that the consequence of this exercise is that all six tosses result in a "head?" Or, in a research setting, we might ask, "Assuming this hypothesis to be true, what is the probability that we get this particular outcome?" Statistics on the other hand deals with the opposite issue: given a particular data set, what can we infer about the possible situations—i.e., theories—that may have generated this situation?

Of critical importance is that probability theory lies at the heart of understanding data analysis. In order to acquire skill and artistry at the art of data analysis, it's necessary to understand probability theory reasonably well. So it is to probability theory that we turn in the next chapter.