1

DESIGNING AND CARRYING OUT A STATISTICAL STUDY

In this chapter we study random behavior and how it can fool us, and we learn how to design studies to gain useful and reliable information. After completing this chapter, you should be able to

- use coin flips to replicate random processes and interpret the results of coin-flipping experiments,
- define and understand probability,
- define, intuitively, *p*-value,
- list the key statistics used in the initial exploration and analysis of data,
- describe the different data formats that you will encounter, including relational database and flat file formats,
- describe the difference between data encountered in traditional statistical research and "big data,"
- explain the use of treatment and control groups in experiments,
- explain the role of randomization in assigning subjects in a study,
- explain the difference between observational studies and experiments.

You may already be familiar with statistics as a method of gathering and reporting data. Sports statistics are a good example of this. For many decades, data have been collected and reported on the performance of both teams and players using standard metrics such as yards via pass completions (quarterbacks in American football), points scored (basketball), and batting average (baseball).

Introductory Statistics and Analytics: A Resampling Perspective, First Edition. Peter C. Bruce.

Sports fans, coaches, analysts, and administrators have a rich array of useful statistics at their disposal, more so than most businesses. TV broadcasters can not only tell you when a professional quarterback's last fumble was but they can also queue up television footage almost instantly, even if that footage dates from the player's college days. To appreciate the role that statistical analysis (also called *data analytics*) plays in the world today, one needs to look no further than the television broadcast of a favorite sport—pay close attention to the statistics that are reported and imagine how they are arrived at.

The whole point in sports, of course, is statistical—to score more points than the other player or the other team. The activities of most businesses and organizations are much more complex, and valid statistical conclusions are more difficult to draw, no matter how much data are available.

Big Data

In most organizations today, raw data are plentiful (often too plentiful), and this is a two-edged sword.

- Huge amounts of data make prediction possible in circumstances where small amounts of data do not help. One type of recommendation system, for example, needs to process large numbers of transactions to locate transactions with the same items you are looking at—enough so that reliable information about associated items can be deduced.
- On the other hand, huge data flows can obscure the signal, and useful data are often difficult and expensive to gather. We need to find ways to get the most information and the most accurate information for each dollar spent in gathering and preparing data.

Data Mining and Data Science

The terms *big data*, *data mining*, *data science*, and *predictive analytics* often go together, and when people think of data mining various things come to mind. Laypersons may think of large corporations or spy agencies combing through petabytes of personal data in hopes of locating tidbits of information that are interesting or useful. Analysts often consider data mining to be much the same as predictive analytics—training statistical models to use known values ("predictor variables") to predict an unknown value of interest (loan default, acceptance of a sales offer, filing a fraudulent insurance claim, or tax return).

In this book, we will focus more on standard research statistics, where data are small and well structured, leaving the mining of larger, more complex data to other books. However, we will offer frequent windows into the world of data science and data mining and point out the connections with the more traditional methods of statistics.

In any case, it is still true that most data science, when it is well practiced, is not just aimless trolling for patterns but starts out with questions of interest such as:

- What additional product should we recommend to a customer?
- Which price will generate more revenue?
- Does the MRI show a malignancy?
- Is a customer likely to terminate a subscription?

All these questions require some understanding of random behavior and all benefit from an understanding of the principles of well-designed statistical studies, so this is where we will start.

1.1 A SMALL EXAMPLE

In the fall of 2009, the Canadian Broadcasting Corporation (CBC) aired a radio news report on a study at a hospital in Quebec. The goal of the study was to reduce medical errors. The hospital instituted a new program in which staff members were encouraged to report any errors they made or saw being made. To accomplish that, the hospital agreed not to punish those who made errors. The news report was very enthusiastic and claimed that medical errors were less than half as common after the new program was begun. An almost parenthetical note at the end of the report mentioned that total errors had not changed much, but major errors had dropped from seven, the year before the plan was begun, to three, the year after (Table 1.1).

TABLE 1.1 Major Errors in a Quebec Hospital

Before no-fault reporting	Seven major errors
After no-fault reporting	Three major errors

1.2 IS CHANCE RESPONSIBLE? THE FOUNDATION OF HYPOTHESIS TESTING

This seems impressive, but a statistician recalling the vitamin E case might wonder if the change is real or if it could just be a fluke of chance. This is a common question in statistics and has been formalized by the practices and policies of two groups:

- Editors of thousands of journals who report the results of scientific research because they want to be sure that the results they publish are real and not chance occurrences.
- Regulatory authorities, mainly in medicine, who want to be sure that the effects of drugs, treatments, and so on are real and are not due to chance.

A standard approach exists for answering the question "is chance responsible?" This approach is called a *hypothesis test*. To conduct one, we first build a plausible mathematical model of what we mean by chance in the situation at hand. Then, we use that model to estimate how likely it is, just by chance, to get a result as impressive as our actual result. If we find that an impressive improvement like the observed outcome would be very unlikely to happen by chance, we are inclined to reject chance as the explanation. If our observed result seems quite possible according to our chance model, we conclude that chance is a reasonable explanation. We now conduct a hypothesis test for the Quebec hospital data.

What do we mean by the outcome being "just" chance? How should that chance model look like? We mean that there is nothing remarkable going on—that is, the no-fault reporting has no effect, and the $7 + 3 = 10$ major errors just happened to land seven in the first

year and three in the second. If there is no treatment effect from no-fault reporting and only chance were operating, we might expect 50/50 or five in each year, but we would not *always* get five each year if the outcome were due to chance. One way that we could see what might happen would be to just toss a coin 10 times, letting the 10 tosses represent the 10 major errors, and letting heads represent the first year and tails the second. Then a toss of HTTHTTHHHH would represent six in the first year and four in the second.

Try It Yourself 1.1

Toss a coin 10 times and record the number of heads and the number of tails. We will call the 10 tosses one trial. Then repeat that trial 11 more times for a total of 12 trials and 120 tosses. To try this exercise on your computer, use the macro-enabled Excel workbook `boxsampler1.xlsm` (located at the book website), which contains a Box Sampler model.

The textbook supplements contain both Resampling Stats for Excel and StatCrunch procedures for this problem.

Did you ever get seven (or more) heads in a trial of 10 tosses? (Answers to "Try it Yourself" exercises are at the end of the chapter.)

Let us recap the building blocks of our model:

- A single coin flip, representing the allocation of a single error to this year (T in the above discussion) or the prior year (H in the above discussion).
- A series of 10 coin flips, representing a single simulation, also called a *trial*, that has the same sample size as the original sample of 10 errors.
- Twelve repetitions of that simulation.

At this stage, you have an initial impression of whether seven or more heads is a rare event. But you only did 12 trials. We picked 12 as an arbitrary number, just to get started. What is next?

One option is to sit down and figure out exactly what the probability is of getting seven heads, eight heads, nine heads, or 10 heads. Recall that our goal is to learn whether seven heads and only three tails are an extreme, that is, it is an unusual occurrence. If we get lots of cases where we get eight heads, nine heads, and so on, then clearly, seven heads is not extreme or unusual.

Why do we count ≥ 7 instead of $=7$? This is an important but often a misunderstood point. If it is not clear, please raise it in class!

We have used the terms "*probability*" and "*chance*," and you probably have a good sense of what they mean, for example, probability of precipitation or chance of precipitation. Still, let us define them—the meaning is the same for each, but probability is a more specific statistical term so we will stick with that.

***Definition:* A somewhat subjective definition of probability**

The probability of something happening is the proportion of time that it is expected to happen when the same process is repeated over and over (paraphrasing from Freedman, et al., *Statistics*, 2nd ed., Norton, 1991, 1st ed. 1978).

***Definition:* Probability defined more like a recipe or formula**

First, turn the problem into a box filled with slips of paper, with each slip representing a possible outcome for an event. For example, a box of airline flights would have a label for each flight: late, on time, or canceled. The probability of an outcome is the number of slips of paper with that outcome divided by the total number of slips of paper in the box.

Question 1.1

From the above, particularly the second definition, you can see that the probability of something happening must always lie between ____ and ____, inclusive.

You can speak in terms of either proportions or percentages—40% is the same as 0.40.

Earlier, we calculated all the possible outcomes for three flips of a coin. Can we do the same thing for 10 flips? If you try it, you will see that this method of counting will quickly become tedious.

Three flips is easier—here is a video from the Khan Academy that illustrates how to calculate the probability of two heads in three tosses by counting up the possibilities.

https://www.youtube.com/watch?v=3UlE8gyKbkU&feature=player_embedded

With 10 flips, one option is to do many more simulations. We will get to that in a bit, but, for now, we will jump to the conclusion so that we can continue with the overall story.

The probability of getting seven or more heads is about $2/12 = 0.1667$.

Interpreting This Result

The value 0.1667 means that such an outcome, i.e., seven or more heads, is not all that unusual, and the results reported from Canada could well be due to chance. This news story was not repeated later that day nor did it appear on the CBC website, so perhaps they heard from a statistician and pulled the story.

Question 1.2

Would you consider chance as a reasonable explanation if there were 10 major errors the year before the change and none the year after? Hint: use the coin tosses that you already performed.

Suppose it had turned out the other way. If our chance model had given a very low probability to the actual outcome, then we are inclined to reject chance as the main factor.

***Definition: p*-value**

If we examine the results of the chance model simulations in this way, the probability of seeing a result as extreme as the observed value is called the *p-value* (or probability value).

Even if our chance model had produced a very low probability, ruling out chance, this does not necessarily mean that the real cause is the new no-fault reporting policy. There are many other possible explanations. Just as we need to rule out chance, we need to rule out those as well. For example, we might be more impressed if our hospital was unique—reducing its errors while every other hospital in Quebec had more major errors the second year. Conversely, we would be less impressed if the number of errors went down at all hospitals that second year—including those with no new program.

Do not worry if this definition of *p*-value and the whole hypothesis testing process are not fully clear to you at this early stage. We will come back to it repeatedly.

The use of *p*-values is widespread; their use as decision-making criteria lies more in the *research* community than in the *data science* community.

Increasing the Sample Size

Intuition tells us that small samples lead to fluke results. Let us see what happens when we increase the sample size.

Try It Yourself 1.2

Let us double the sample size and imagine that the study had revealed 14 errors in 1 year and six the following, instead of seven and three. Now, regroup your 12 simulations of 10 tosses into six trials of 20 tosses each. Combine the first and the second sets, the third and fourth, and so on. Then do six more trials of 20 tosses each for a total of 120 additional tosses. You should now have 12 sets of 20 tosses. If you want to try a computer simulation, use the Box Sampler macro-enabled Excel workbook boxsampler2.xlsm.

The textbook supplements contain a Resampling Stats procedure for this problem. Did you ever get 14 or more heads in a trial of 20 tosses?

Technique

We will use the "Technique" heading for the details you need to do the analyses. We illustrate the use of a computer to generate random numbers, which is shown as follows.

In our original example, we saw seven errors in the first year and three errors in the next, for a *reduction* of four errors. As we develop this example further, we will deal exclusively with data on *reduction* in errors.

Tossing coins can get tiresome and can only model events that have a 50/50 chance of either happening or not happening. Modeling random events is typically done by generating random numbers by computer.

Excel, for example, has two options for generating random numbers:

RAND generates a random number between 0 and 1.

RANDBETWEEN generates a random integer between two values that you specify.

You then need to map the random digit that is generated to the outcome of an event that you are trying to model. For example:

1. A customer of a music streaming subscription service has a 15% probability of canceling the service in a given year. This could be modeled by generating a random integer between 1 and 100 and labeling 1–15 as "cancel" and 16–100 as "maintain subscription."

In Excel, the function would be entered as =RANDBETWEEN(1,100).

After generating, say, 1000 random numbers (and putting them in cells A1:A1000), you could count the number of cancelations using COUNTIF:

=COUNTIF(A1:A1000,"<=15").

What is a Random Number?

For our purposes, we can think of a random number as the result of placing the digits 0–9 in a hat or a box, shuffling the hat or box, and then drawing a digit. Most random numbers are produced by computer algorithms that produce series of numbers that are effectively random and unpredictable, or at least sufficiently random for the purpose at hand. But the numbers are produced by an algorithm that is technically called a *pseudo-random number generator*. There have been many research studies and scholarly publications on the properties of random number generators (RNGs) and the computer algorithms they use to produce pseudo random numbers. Some are better than others; the details of how they work are beyond the scope of this book. We can simply think of random number generators as the computer equivalent of picking cards from a hat or a box that has been well shuffled.

1.3 A MAJOR EXAMPLE

To tie together our study of statistics, we will look at one major example. Using the study reported by the CBC as our starting point, we introduce basic but important statistical concepts.

Imagine that you have just been asked to design a better study to determine if the sort of no-fault accident reporting tried in a Quebec hospital really does reduce the number of serious medical errors. The standard type of study in such a situation would be an *experiment*.

Experiment versus Observational Study

In the fifth inning of the third game of the 1932 baseball World Series between the NY Yankees and the Chicago Cubs, the great slugger Babe Ruth came to bat and pointed toward center field as if to indicate that he planned to hit the next pitch there. On the next pitch, he indeed hit the ball for a home run into the centerfield bleachers.*

A Babe Ruth home run was an impressive feat but not that uncommon. He hit one every 11.8 at bats. What made this one so special is that he predicted it. In statistical terms, he specified in advance a theory about a future event—the next swing of the bat—and an outcome of interest—home run to centerfield.

In statistics, we make an important distinction between studying preexisting data—an observational study—and collecting data to answer a prespecified question—an experiment or a prospective study.

We will learn more about this later but keep in mind that the most impressive and durable results in science come when the researcher specifies a question in advance and then collects data in a well-designed experiment to answer the question. Offering commentary on the past can be helpful but is no match for predicting the future.

*There is some controversy about whether he actually pointed to center field or to left field and whether he was foreshadowing a prospective home run or taunting Cubs players. You can Google the incident ("Babe Ruth called shot") and study videos on YouTube and then judge for yourself.

1.4 DESIGNING AN EXPERIMENT

The principles of designing an experiment are fundamental and should be studied by both data scientists and research statisticians. When it comes to practice and implementation, this material will be of primary interest to the *research* community identified at the beginning of the introduction.

In our errors experiment, we could compare two groups of hospitals. One group uses the no-fault plan and the other does not. The group that gets the change in treatment you wish to study is called the *treatment group*. The group that gets no treatment or the standard treatment is called the *control group*. Normally, you need some reference group for comparison, although in some studies you may be comparing multiple treatments with no control. How do you decide which hospitals go into which group?

You would like the two groups to be similar to one another, except for the treatment/control difference. That way, if the treatment group does turn out to have fewer errors, you can be confident that it was due to the treatment. One way to do this would be to study all the hospitals in detail, examine all their relevant characteristics, and assign them to treatment/control in such a way that the two groups end up being similar across all these attributes. There are two problems with this approach.

1. It is usually not possible to think of all the relevant characteristics that might affect the outcome. Research is replete with the discovery of factors that were unknown prior to the study or thought to be unimportant.
2. The researcher, who has a stake in the outcome of the experiment, may consciously or unconsciously assign hospitals in a way that enhances the chances of the success of his or her pet theory.

Oddly enough, the best strategy is to assign hospitals randomly—perhaps by tossing a coin.

Randomizing

True random assignment eliminates both conscious and unconscious bias in the assignment to groups. It does not guarantee that the groups will be equal in all respects. However, it does guarantee that any departure from equality will be due simply to the chance allocation and that the larger the number of samples, the fewer differences the groups will have. With extremely large samples, differences due to chance virtually disappear and you are left with differences that are real—provided the assignment to groups is really random.

Law of Large Numbers

The law of large numbers states that, despite short-term average deviations from an event's theoretical mean, such as the chance of a coin landing heads, the long-run empirical—actual—average occurrence of the event will approach, with greater and greater precision, the theoretical mean. The short-run deviations get washed out in a flood of trials. During World War II, John Kerrich, a South African mathematician, was imprisoned in Denmark. In his idle moments, he conducted several probability experiments.

In one such experiment, he flipped a coin repeatedly, keeping track of the number of flips and the number of heads. After 20 flips, he was exactly even—10 heads and 10 tails. After 100 flips, he was down six heads—44 heads and 56 tails—or 6%. After 500 flips, he was up five heads—255 heads and 245 tails—or 1%. After 10,000 flips, he was up 67 heads or 0.67%.

A plot of all his results with the proportion of heads on the *y*-axis and the number of tosses on the *x*-axis shows a line that bounces around a lot on the left side but settles down to a straighter and straighter line on the right side, tending toward 50%.

Do not confuse the Law of Large Numbers with the popular conception of the Law of Averages.

Law of Large Numbers

Long run actual average will approach the theoretical average.

Law of Averages

A vague term, sometimes meaning as mentioned earlier but also used popularly to refer to the mistaken belief that, after a string of heads, the coin is "due" to land tails, thereby preserving its 50/50 probability in the long run. One often encounters this concept in sports, for example, a batter is "due" for a hit after a dry spell.

Random assignment let us make the claim that any difference in the group outcomes that can *more* than might happen by chance is, in fact, due to the different treatment received by the groups. Kerrich had a lot of time on his hands and could accumulate a huge sample under controlled conditions for his simple problem. In actual studies, researchers rarely have the ability to collect samples sufficiently large that we can dismiss chance as a factor. In the study of probability in this course, lets us quantify the role that chance can play and take it into account (Figure 1.1).

Even if we performed a dummy experiment in which both groups got the same treatment, we would expect to see some differences from one hospital to another. An everyday example of this might be tossing a coin. You get different results from one toss to the next just by chance. Check the coin tosses you did earlier in connection with the CBC news report on

Figure 1.1 Kerrich coin tosses. Number of tosses on the *x*-axis and the proportion of heads on the *y*-axis.

medical errors. If you toss a coin 10 times, you will get a certain number of heads. Do it again and you will probably get a different number of heads.

Although the results vary, there are laws of chance that allow you to calculate things like how many heads you would expect on average or how much the results would vary from one set of 10 tosses to the next. If we assign subjects at random, we can use these same laws of chance—or a lot of coin tosses—to analyze our results.

If we have Doctor Jones assign subjects using her own best judgment, we will have no mathematical theory to guide us. That is because it is very unlikely that we can find any books on how Doctor Jones assigns hospitals to the treatment and control groups. However, we can find many books on random assignment. It is a standard, objective way of doing things that works the same for everybody. Unfortunately, it is not always possible. Human subjects can neither be assigned a gender nor a disease.

Planning

You need some hospitals and you estimate that you can find about 100 within reasonable distance. You will probably need to present a plan for your study to the hospitals to get their approval. This seems like a nuisance, but they cannot let just anyone do any study they please on the patients. Studies of new prescription drugs require government approval as well, which is a long and costly process. In addition to writing a plan to get approval, you know that one of the biggest problems in interpreting studies is that many are poorly designed. You want to avoid that so you think carefully about your plan and ask others for advice. It would be good to talk to a statistician who has experience in medical work. Your plan is to ask the 100 or so available hospitals if they are willing to join your study. They have the right to say no. You hope that quite a few will say yes. In particular, you hope to recruit 50 willing hospitals and randomly assign them to two groups of 25.

> ***Try It Yourself 1.3***
>
> Suppose you wanted to study the impact of watching television on violent behavior with an experiment. What issues might you encounter in trying to assign treatments to subjects? What would the treatment be?

Blinding

We saw that randomization is used to try to make the two groups similar at the beginning. It is important to keep them as similar as possible. We want to be sure that the treatment is the only difference between them. One subtle difference we have to worry about when working with humans is that their behavior can be changed by the fact that they are participating in a study.

Out-of-Control Toyotas?

In the fall of 2009, the National Highway Transportation Safety Agency received several dozen complaints per month about Toyota cars speeding out of control. The rate of complaint was not that different from the rates of complaint for other car companies. Then, in November of 2009, Toyota recalled 3.8 million vehicles to check for sticking gas pedals. By February, the complaint rate had risen from several dozen per month to over 1500 per month of alleged cases of unintended acceleration. Attention turned to the electronic throttle.

Clearly, what changed was not the actual condition of cars—the stock of Toyotas on the road in February of 2010 was not that different from November of 2009. What changed was car owners' awareness and perception as a result of the headlines surrounding the recall. Acceleration problems, whether real or illusory, that escaped notice before November 2009 became causes for worry and a trip to the dealer. Later, the NHTSA examined a number of engine data recorders from accidents where the driver claimed to have experienced acceleration despite applying the brakes. In all cases, the data recorders showed that the brakes were not applied.

In February 2011, the US Department of Transportation announced that a 10-month investigation of the electronic throttle showed no problems.

In April 2011, a jury in Islip, NY took less than an hour to reject a driver's claim that a mispositioned floor mat caused his Toyota to accelerate and crash into a tree. The jury's verdict? Driver error.

As of this writing, we still do not know the actual extent of the problem. But from the evidence to date, it is clear that public awareness of the problem boosted the rate of complaint far out of proportion to its true scope.

Lesson: Your perception of whether you personally experience a problem or benefit is substantially affected by your prior awareness of others' problems/benefits.

Sources: *Wall Street Journal*, July 14, 2010; The Analysis Group (http://www.analysisgroup.com/auto_safety_analysis.aspx—accessed July 14, 2010); *A Today* online, April 2, 2011.

In some situations, we can avoid telling people that they are participating in a study. For example, a marketing study might try different ads or products in various regions without publicizing that they are doing so for research purposes. In other situations, we may not be able to avoid letting subjects know they are being studied, but we may be able to conceal whether they are in the treatment or control group. One way is to impose a dummy treatment on the control group.

Such a dummy treatment is called a *placebo*. It is especially important when we decide how well the treatment worked by asking the subjects. Experience has shown that subjects will often report good results even for dummy treatments. Part of this is that people want

to please, or at least not offend, the researcher. Another part is that people may believe in the treatment and therefore think that it helped even when it did not. The researcher may communicate this positive expectation. For this reason, we prefer that neither the subjects nor any researchers in contact with the subjects know whether the subjects are getting the real treatment or the placebo. Then we hope that the researchers will communicate identical expectations to both groups, and the subjects will be equally eager to please or to expect equally good results. Experience has also shown that people respond positively to attention and just being part of a study may cause subjects to improve. This positive response to the knowledge that you are being treated is called the *placebo effect*. More specifically, positive response to the attention of participating in a study is called the *Hawthorne effect.*

We say a study is *single-blind* when the subjects—the hospitals in our medical errors example—do not know whether they are getting the treatment. It is *double-blind* if the staff in contact with the subjects also does not know. It is *triple-blind* if the people who evaluate the results do not know either. These people might be lab technicians who perform lab tests for the subjects but never meet them. They cannot communicate any expectations to the subjects, but they may be biased in favor of the treatment when they perform the tests.

It is not always practical to have all these levels of blinding. A reasonable compromise might be necessary. For our hypothetical study of medical errors, we cannot prevent the hospitals from knowing that they are being studied because we need their agreement to participate. It may be unethical to have the control group do *nothing* to reduce medical errors. What we might be able to do is consult current practices on methods for reducing medical errors and codify them. Then ask the treatment hospitals to implement those best practices PLUS no-fault-reporting and those at the control hospitals to simply implement the basic best practices code. This way, all hospitals receive a treatment but do not know which one is of actual interest to the researcher.

Try It Yourself 1.4

How would you use blinding in a study to assess the effects of watching television on violent behavior?

In addition to the various forms of blinding, we try to keep all other aspects of each subject's environment the same. This usually requires spelling out in great detail what will be done. For example, when no experiment is being conducted, each individual hospital might decide on how to deal with medical errors. In an experiment, we need to agree on a common method that will be applied to all hospitals. We would try to find hospitals that are as similar as possible and maintain patient conditions as similarly as possible. By keeping the two groups the same in every way except the treatment, we can be confident that any differences in the results were due to it. Any difference in the outcome due to nonrandom extraneous factors is called *bias*. Statistical bias is not the same as the type that refers to people's opinions or states of mind.

Try It Yourself 1.5

What factors other than watching television might affect violent behavior? How would you control these in a study to assess the effects of watching television on violent behavior?

Before-After Pairing

We could run our study for a year and measure the total number of medical errors each hospital had by the end of that period. A better strategy is to measure how many errors they had the year before the study as well. Then, we have *paired data*—two measurements on each unit. This allows us to compare the treatment to no treatment *at the same hospitals*. The study reported by the CBC had both before and after data on the same hospital.

Even if we use before/after measurements on the same hospitals, we should also retain the control group. Having both a control group and a treatment group allows us to separate out the improvement due to no-fault-reporting from the improvement due to the more general best practices treatment. Having a control group also controls for trends that affect *all* hospitals. For example, the number of errors could be increasing due to an increased patient load at hospitals, generally, or decreasing due to better doctor training or greater awareness of the issue—perhaps generated by CBC news coverage. The vitamin E study compared two groups over the same time period but did not have before and after data.

> ***Try It Yourself 1.6***
>
> How could you use a control group or pairing in a study to assess the effects of watching television on violent behavior?

1.5 WHAT TO MEASURE—CENTRAL LOCATION

Part of the plan for any experiment will be the choice of what to measure to see if the treatment works. This is a good place to review the standard measures with which statisticians are concerned: central location of and variation in the data.

Mean

The *mean* is the average value—the sum of all the values divided by the number of values. It is generally what we use unless we have some reason not to use it.

Consider the following set of numbers: {3 5 1 2}
The mean is $(3 + 5 + 1 + 2)/4 = 11/4 = 2.75$.
You will encounter the following symbols for the mean:
$\bar{x}$ represents the mean of a *sample* from a population. It is written as *x*-bar in inline text.
μ represents the mean of a *population*. The symbol is the Greek letter mu.

Why make the distinction? Information about samples is observed, and information about large populations is often inferred from smaller samples. Statisticians like to keep the two things separate in the symbology.

Median

The *median* is the middle number on a sorted list of the data. Table 1.2 shows the sorted data for both groups in the hospital.

The middle number on each list would be the 13th value (12 above and 12 below). If there is an even number of data values, the middle value is one that is not actually in the data set but rather is the average of the two values that divide the sorted data into upper and lower halves.

TABLE 1.2 Hospital Error Reductions, Treatment, and Control Groups

Control	Treatment
1	2
1	2
1	2
1	2
1	2
1	2
1	2
1	2
1	2
1	2
1	2
1	2
2	2
2	2
2	2
2	2
2	2
2	2
2	3
2	3
3	4
3	4
4	5
4	6
5	9

We find that the median is the same for both lists! It is 2. This is not unusual for data with a lot of repeated values. The median is a blunt instrument for describing such data. From what we have seen so far, the groups seem to be different. The median does not capture that. Looking at the numbers, you can see the problem. In the control group, the numbers coming before the 2 at Position 13 are all ones; for the treatment group they are all 2s. The median reflects what is happening at the center of the sorted data but not what is happening before or after the center.

The median is more typically used for data measured over a broad range where we want to get an idea of the typical case without letting extreme cases skew the results. Let us say we want to look at typical household incomes in the neighborhoods around Lake Washington in Seattle. In comparing the Medina neighborhood to the Windermere neighborhood, using the mean would produce very different results because Bill Gates lives in Medina. If we use the median, it will not matter how rich Bill Gates is—the position of the middle observation will remain the same.

Question 1.3

A student gave seven as the median of the numbers 3, 9, 7, 4, 5. What do you think he or she did wrong?

Mode

The *mode* is the value that appears most often in the data, assuming there is such a value. In most parts of the United States, the mode for religious preference would be Christian. For our data on errors, the mode is 2 for all 50 subjects and 1 for the control group. The mode is the only simple summary statistic for categorical data, and it is widely used for that. At different times in the history of the United States, the mode for the make of new cars sold each year has been Buick, Ford, Chevrolet, and Toyota. The mode is rarely used for measurement data.

Expected Value

The expected value is calculated as follows.

1. Multiply each outcome by its probability of occurring.
2. Sum these values.

For example, suppose that a local charitable organization organizes a game in which contestants purchase the right to spin a giant wheel with 50 equal-sized sections and an indicator that points to a section when the wheel stops spinning. The right to spin the wheel costs \$5 per spin. One section is marked \$50—that is how much the purchaser wins if the spinner ends up on that section. Five sections are marked \$15, 10 sections are marked \$5, and the remaining sections are marked \$0.

To calculate the expected value of a spin, the outcomes, with the purchase price of the spin subtracted from the prize, are multiplied by their probabilities and then summed.

$$\text{EV} = \left(\frac{1}{50}\right)(\$50 - \$5) + \left(\frac{5}{50}\right)(\$15 - \$5) + \left(\frac{10}{50}\right)(\$5 - \$5) + \left(\frac{34}{50}\right)(\$0 - \$5)$$

$$\text{EV} = -\$1.50$$

The expected value favors the charitable organization, as it probably should. For each ticket you purchase, you can expect to lose, on average, \$1.50. Of course, you will not lose exactly \$1.50 in any of the above scenarios. Rather, the \$1.50 is what you would lose per ticket, on average, if you kept playing this game indefinitely.

The expected value is really a fancier mean; it adds the ideas of future expectations and probability weights. Expected value is a fundamental concept in business valuation and capital budgeting—the expected number of barrels of oil a new well might produce, for example, the expected value of 5 years of profit from new acquisition or the expected cost savings from new patient management software at a clinic.

Percents

Percents are simply proportions multiplied by 100. Percents are often used in reporting as they can be understood and visualized a bit more easily and intuitively than proportions.

Proportions for Binary Data

Definition: **Binary data**

Binary data is data that can take one of only two possible outcomes—win/lose, survive/die, purchase/do not purchase.

When you have binary data, the measure of central tendency is the proportion. An example would be the proportion of the survey approving of the president. The proportion for binary data fully defines the data—once you know the proportion, you know all the values. For example, if you have a sample of 50 zeros and ones, and the proportion for one is 60%, then you know that there are 30 ones and 20 zeros.

For the convenience of software and analysis, binary data are often represented as 0s and 1s. For purely arbitrary reasons, a "1" is called a *success*, but this term has no normative meaning and simply indicates the outcome associated with some action or event of interest. For example, in a data set used to analyze college dropouts, a "1" might be used to indicate dropout. With binary data in which one class is much more scarce than the other (e.g., fraud/no-fraud or dropout/no-dropout), the scarce class is often designated as "1."

1.6 WHAT TO MEASURE—VARIABILITY

If all the hospitals in the control group had one fewer error and all those in the treatment group had two fewer, our job would be easy. We would be very confident that the treatment improved the reduction in the number of errors by exactly one. Instead, we have a lot of variability in both batches of numbers. This just means that they are not all the same.

Variability lies at the heart of statistics: measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability, and making decisions in the presence of it.

Just as there are different ways to measure central tendency—mean, median, mode—there are also different ways to measure variability.

Range

The *range* of a batch of numbers is the difference between the largest and smallest number. Referring to Table 1.2, the range for the control group is $5 - 1 = 4$. Note that in statistics the range is a single number.

> ***Try It Yourself 1.7***
>
> Referring to the same table, what is the range for the treatment group?

The range is very sensitive to outliers. Recall the two similar Seattle neighborhoods—Windermere and Medina. The range of income in Medina, where Bill Gates lives, will be much larger than the range in Windermere.

Percentiles

One way to get around the sensitivity of the range to outliers is to go in a bit from each end and take the difference from there. For example, we could take the range between the 10th percentile and the 90th percentile. This would eliminate the influence of extreme observations.

Definition: **Pth percentile**

In a population or a sample, the Pth percentile is a value such that at least P percent of the values take on this value or less and at least (100 − P) percent of the values take on this value or more. Sometimes, there is a single value in the data that satisfies this requirement and sometimes there are two. In the latter case, it is best to take the midpoint between the two values that do. Software may have slightly differing approaches that can produce differing answers.

More intuitively: to find the 80th percentile, sort the data. Then, starting with the smallest value, proceed 80% of the way to the largest value.

Interquartile Range

One common approach is to take the difference between the 25th percentile and the 75th percentile.

Definition: **Interquartile range**

The interquartile range (or IQR) is the 75th percentile value minus the 25th percentile value. The 25th percentile is the first quartile, the 50th percentile is the second quartile, also called the *median*, and the 75th percentile is the third quartile. The 25th and 75th percentiles are also called *hinges*.

Here is a simple example: 3, 1, 5, 3, 6, 7, 2, 9. We sort these to get 1, 2, 3, 3, 5, 6, 7, 9. The 25th percentile is at 2.5 and the 75th percentile is at 6.5, so the interquartile range is $6.5 - 2.5 = 4$. Again, software can have slightly differing approaches that yield different answers.

> ***Try It Yourself 1.8***
>
> Find the IQR for the control data, the treatment data, and for all 50 observations combined.

Deviations and Residuals

There are also a number of measures of variability based on deviations from some typical value. Such deviations are called *residuals*.

Definition: **Residual**

A residual is the difference between a mean value and an observed value or the difference between a value predicted by a statistical model and an actual observed value.

For 1, 4, 4, the mean is 3 and the median is 4. The deviations from the mean are the differences

$$1 - 3 = -2 \qquad 4 - 3 = 1 \qquad 4 - 3 = 1$$

Mean Absolute Deviation

One way to measure variability is to take some kind of typical value for these residuals. We could take the absolute values of the deviations—{2 1 1} in the above case and then average them: $(2 + 1 + 1)/3 = 1.33$. Taking the deviations themselves, without taking the absolute values, would not tell us much—the negative deviations exactly offset the positive ones. This always happens with the mean.

Variance and Standard Deviation

Another way to deal with the problem of positive residuals offsetting negative ones is by squaring the residuals.

Definition: **Variance for a population**

The *variance* is the mean of the squared residuals, where $\mu =$ population mean, x represents the individual population values, and $N =$ population size.

$$\text{Variance} = \sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

The *standard deviation* σ is the square root of the variance. The symbol σ is the Greek letter *sigma* and commonly denotes the standard deviation.

The appropriate Excel functions are VARP and STDEVP. The P in these functions indicates that the metric is appropriate for use where the data range is the entire population being investigated; that is, the study group is not a sample.

The standard deviation is a fairly universal measure of variability in statistics for two reasons: (i) it measures typical variation in the same units and scale as the original data and (ii) it is mathematically convenient, as squares and square roots can effectively be plugged into more complex formulas. Absolute values encounter problems on that front.

> ***Try It Yourself 1.9***
>
> Find the variance and standard deviation of 8, 1, 4, 2, 5 by hand. Is the standard deviation in the ballpark of the residuals, that is, the same order of magnitude?

Variance and Standard Deviation for a Sample

When we look at a sample of data taken from a larger population, we usually want the variance and, especially, the standard deviation—not in their own right but as estimates of these values in the larger populations.

Intuitively, we are tempted to estimate a population metric by using the same metric in the sample. For example, we can estimate the population mean effectively by using the sample mean or the population proportion using the sample proportion.

The same is not true for measures of variability. The range in a sample (particularly a small one) is almost always going to be biased—it will usually be less than the range for the population.

Likewise, the sample variance and standard deviation are *not* the best estimates for the population values because they consistently underestimate the variance and standard deviations in the population being sampled.

However, if you divide by $n-1$ instead of n, the variance and standard deviation from a sample become unbiased estimators of the population values. A mathematical proof is beyond the scope of this course, but you can demonstrate this fact with the "Try It Yourself" exercise below.

Definition: **Sample variance**

$$\text{Sample variance} = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Definition: **Sample standard deviation**

The sample standard deviation s is the square root of the sample variance.

The appropriate Excel functions are VAR and STDEV.

In statistics, you will encounter the term degrees of freedom. Its exact definition is not needed for this course, but the concept can be illustrated here. Let us say you have three observations and you know that their variance is x. Once you know the first two values, the third is predetermined by the first two and the value for the variance. We say there are $n - 1$, in this case, two, degrees of freedom. The denominator in the sample variance formula is the number of degrees of freedom.

Try It Yourself 1.10

In your resampling software, randomly generate a population of 1000 values. It does not matter what population you generate—let us say a population of randomly selected numbers between 0 and 9. In Excel, you can do this with the RANDBETWEEN function. Next, find the variance of this population using the population variance formula. Then, repeatedly take resamples of size 10 and calculate the variance for each resample according to the same population formula. How does the mean of the resample variances compare to the population variance?

Tutorials for this exercise using Resampling Stats for Excel and StatCrunch can be found in the textbook supplements.

For a Box Sampler resampling tutorial based on this exercise, see the file box_sampler_tutorial.pdf.

1.7 WHAT TO MEASURE—DISTANCE (NEARNESS)

The concept of statistical distance is of particular interest to the *data science* community identified at the beginning of the Introduction.

Consider a poll in which respondents are asked to assess their preferences for the musical genres listed below. Ratings are on a scale of 1 (dislike) to 10 (like) and we have poll results from three students (Table 1.3).

TABLE 1.3 Musical Genre Preferences

Person	Rock	Rap	Country	Jazz	New Age
A	7	1	9	1	3
B	4	9	1	3	1
C	9	1	7	2	2

Consider person C. Is she more like person A or person B? Looking at the scores, our guess would be that person C is more like person A. We can measure this distance statistically by subtracting one vector from the another, squaring the differences so they are all positive, summing them so we have a single number, then taking the square root so the original scale is restored.

Definition: **Vector**

A vector is a row of numbers. Vector arithmetic is done by performing the operation on the corresponding elements of each vector, resulting in a new vector of sums, differences, products, and so on.

The statistic that we have described is "Euclidean Distance." Here is the formula, followed by the calculations.

As a general example, assume that we have two vectors, w and x, each containing n values. The Euclidean distance between the two vectors is

$$\text{Euclidean Distance} = \sqrt{(w1 - x1)^2 + (w2 - x2)^2 + (w3 - x3)^2 + \cdots + (wn - xn)^2}$$

If you look carefully at the formula, you might recognize that this is the general formula for the distance between two points from high school geometry. The formula mentioned earlier is for n dimensions, whereas high school math courses usually work with two or three dimensions.

For a specific example, the Euclidean Distance between vectors A and B—a measure of how alike person A is to person B—in the table shown earlier is calculated (Table 1.4).

In the table shown earlier, the sum of the squares of the differences of each row is 145. The square root of 145 is 12.04, which is the Euclidean Distance between the two vectors representing person A and person B. Looking back at the data in Table 1.3, try the following problem:

Question 1.4

Let us say that you own a music store. A, B, and C are all customers of yours, and A and B have both just made purchases. You want to recommend one of these purchases

TABLE 1.4 Euclidean Distance Between A and B

	Rock	Rap	Country	Jazz	New Age
Person A	7	1	9	1	3
Person B	4	9	1	3	1
$(A-B)^2$	9	64	64	4	4
Sum $(A-B)^2$	145				
Euclidean distance	12.04				

for C. Which one would you recommend? See the file euclidean_distance.xls for the answer.

Distance measures are used in statistics for multiple purposes:

- Finding clusters or segments of customers who are like one another.
- Classifying records by assigning them the same class as nearby records.
- Locating outliers, for example, airport security screening.
- Finding the distance to a benchmark. For example, if you have a list of symptoms for an individual, what disease is it closest to?

1.8 TEST STATISTIC

Let us continue with our analysis of the hospital data, using the means for error reduction. Certainly, 2.80 is bigger than 1.88. How much bigger? We could say it is $2.80 - 1.88 = 0.92$ bigger. That is, to say that the treatment seems to reduce the number of errors by nearly one. But there are other ways to look at this. The ratio $2.80/1.88 = 1.49$ gives another comparison. It says the reduction in errors for the treatment group is 1.49 times that in the control group or nearly 50% greater.

Just counting errors treats them all alike. Ideally, we like to have some sort of measurement along a scale. For example, we assign a number of points to indicate the level of severity for each error and we total the points for each hospital. That would require training someone to assign points in a consistent manner from case to case. This could be expensive and wasteful if the person spends most of their time on minor errors that do not actually impact patient health.

As a compromise, our researcher decides to count the number of "major" errors and use that as a measure. The hospitals will be given simple criteria that will allow errors that might be considered major to be identified. Then, a trained expert will make the final decision as to which errors are major. To make sure that there are no differences in how the decision is made from hospital to hospital, the researcher asks that one expert to do all the counting. If possible, relevant records will be submitted anonymously to the expert so that he or she does not know which hospital the records came from or whether that hospital is in the treatment or control group.

Try It Yourself 1.11

In studying the impact of watching television on violent behavior, how would you *measure* television watching and violent behavior in an assessment study? Would you consider an hour of watching Rocky and Bullwinkle reruns to be equivalent to watching an hour of live coverage of the war in Afghanistan or a boxing match? What violence rating would you give to robbing a convenience store, becoming a professional boxer, or joining the army to fight in an active combat theater?

Our test statistic will be calculated as follows.

1. Measure the number of major medical errors for each hospital for the year before and the year after the treatment is initiated and find the reduction: errors before minus errors after.
2. Calculate the mean reduction for the control group and the treatment group.
3. Find the difference: treatment minus control = 0.92.

Important: Throughout this example, we will be talking about "*reductions* in number of errors," not in the number of errors.

A test statistic is the key measurement that we will use to judge the results of the experiment or study.

Test Statistic for This Study

Mean reduction in errors (treatment) minus mean reduction in errors (control).

1.9 THE DATA

After performing the study as planned, the researchers will need to enter the data into a computer and proofread it for errors. After doing that, they obtained the results as shown in Table 1.5.

TABLE 1.5 Reduction in Major Errors in Hospitals (Hypothetical Extension of the Earlier Example)

Row	Hospital#	Treat?	Reduction in Errors
1	239	0	3
2	1126	0	1
3	1161	0	2
4	1293	1	2
5	1462	1	2
6	1486	0	2
7	1698	1	5
8	1710	0	1
9	1807	0	1
10	1936	1	2
11	1965	1	2
12	2021	1	2

TABLE 1.5 (*Continued*)

Row	Hospital#	Treat?	Reduction in Errors
13	2026	0	1
14	202	0	3
15	208	1	4
16	2269	1	2
17	2381	1	2
18	2388	0	1
19	2400	1	2
20	2475	0	4
21	2548	0	1
22	2551	0	2
23	2661	0	1
24	2677	1	4
25	2739	1	2
26	2795	1	3
27	2889	0	5
28	2892	1	9
29	2991	1	2
30	3166	1	2
31	3190	0	1
32	3254	0	4
33	3312	1	2
34	3373	1	2
35	3403	1	3
36	3403	0	1
37	3429	1	2
38	3441	1	6
39	3520	0	1
40	3568	1	2
41	3580	0	2
42	3599	0	2
43	3660	1	2
44	3985	0	2
45	4014	1	2
46	4060	0	1
47	4076	1	2
48	4093	0	1
49	4230	0	2
50	5633	0	2

Remember, we are counting not the number of errors per hospital but rather the *reduction* in errors.

Database Format

This is a standard database format, which all database programs and most standard-purpose statistical software programs use. The rows represent records or cases—hospitals in this

example and the columns represent variables, which are data that change from hospital to hospital. The format has two key features, which is required by most statistical software.

1. Each row contains all the information for one and only one hospital.
2. All data for a given variable are in a single column.

Spreadsheets like Excel can deal with data equally as rows or columns, but statistical software expects rows to be records and columns to be variables.

Technique

Although it is possible to enter data into Excel in the above format, not all statistical analyses in Excel can cope with having "group" as a variable. Some procedures want the observations arranged in columns according to which group they are in.

Let us look at the parts. The first column is simply the row number.

Column 2, hospital, contains *case labels*. These are arbitrary labels for the experimental units—a unique number for each unit. Case labels keep track of the data. For example, if we find a mistake in the data, we would need to know which hospital that came from so we could investigate the cause and correct the mistake. Numerical codes are preferred to more informative labels when we wish to conceal the group to which subjects were assigned.

Column 3 labels observations from the treatment group with a one and those from the control group with a zero.

Column 4 is the number of major medical errors at the year before the study minus the number from the following year. A positive number represents a reduction in medical errors. Note that all the numbers are positive—things got better whether subjects got the treatment or not! This could be due to the Hawthorne effect or any extra care the subjects got from being in the experiment or due to any number of other factors that may have changed between the 2 years.

Relational Database Format

Most statistical procedures work with data that are in the format as mentioned earlier—a single table in which columns have variables and rows have individual records. And most statistics courses considerately provide data for students in this format. However, this is not how most organizations store and use data.

The ability to extract data from relational databases for analysis will be of particular importance to those in the *data science* community identified in the Introduction.

Consider a hypothetical jobs database and start with the following information:

Steve Walters, a scientist with 12 years experience who lives in Palo Alto, CA, has applied for a position at HSBC in London as a Data Scientist; there are two such positions with HSBC. Data munging (the ability to extract data and prepare it for analysis) is a required skill. This information might be presented in a single record as follows:

Candidate	Skill Set	Exp	Home	Company	Location	No. of Positions	Position	Skill
S Walters	Scientist	12	Palo Alto	HSBC	London	2	Data Scientist	Munging

Consider also that James Morgan, a banker with 17 years of experience who lives in NY, is also being considered for a position with HSBC in London, but the position title for his job is Banker and the skill required is fixed equity knowledge. This record might look like this:

Candidate	Skill Set	Exp	Home	Company	Location	No. of Positions	Position	Skill
J Morgan	Banker	17	NY	HSBC	London	2	Banker	Fixed equity

When combined, the database now looks like this:

Candidate	Skill Set	Exp	Home	Company	Location	No. of Positions	Position	Skill
S Walters	Scientist	12	Palo Alto	HSBC	London	2	Data Scientist	Munging
J Morgan	Banker	17	NY	HSBC	London	2	Banker	Fixed equity

This table format is known as a "*flat file*."

Definition: **Flat file**

A flat file is a table that has two dimensions—rows and columns.

Note the redundancy in the columns for company, location, and number of positions. This duplication will be multiplied as we consider more candidates, more companies, and more jobs. In a customer database, for example, hundreds of invoices might be linked to a single customer. It would be nice to have a structure that allowed for a single table of customers (where all their address, demographic, and contact information is stored) and a separate table for invoices, with each invoice linked to a customer by a single customer number. Structured information like this is usually stored not in flat files but in relational databases.

Definition: **Relational database**

A relational database is composed as a set of tables, each of which has a key column used to relate the information in one table to another.

Definition: **Database normalization**

Normalization of a database is the process of organizing data so that it is stored in a set of related tables with defined linkages. Be sure to distinguish this definition of normalization from the statistical term.

For example, the above information might be stored in three separate tables—one for the candidates, one for the employers, and one for the positions (Tables 1.6–1.8).

The left column in each table is a key, used to connect one table to another. Consider the following table, which uses the keys to establish the relationship among these tables (Table 1.9).

We interpret the first row as follows:

Candidate c1, S. Walters (the scientist from Palo Alto with 12 years of experience), has applied for a data scientist job (munging is the required skill) with HSBC in London, which

TABLE 1.6 Candidate Table

Cno	Candidate	Skill Set	Exp	Home
c1	S Walters	Scientist	12	Palo Alto
c2	J Morgan	Banker	17	NY
c3	W Weingart	Graphic designer	17	Berlin
c4	D Hvorostovsky	Baritone	19	London

TABLE 1.7 Employer Table

Eno	Company	No. of Positions	Location
e1	HSBC	2	London
e2	Twitter	10	NY
e3	Royal Opera	3	London
e4	Google	3	Palo Alto

TABLE 1.8 Job Table

Jno	Position	Specialty
J1	Data scientist	Munging
J2	Banker	Fixed equity
J3	Attorney	bankruptcy
J4	Graphic designer	3-D Animation

TABLE 1.9 Relationship Table

Eno	Jno	Cno	Startdate	Rate
e1	j1	c1	November 13, 2013	125
e1	j2	c2	November 25, 2013	220
e3	j3	c4	January 12, 2014	180
etc				

has two positions open. For the particular job he is applying for, the start data is November 13, 2013 and the pay rate is $125,000 per year.

Organizing the data like this reduces duplication but also allows us to query the database in a structured manner and efficiently extract information.

Definition: **Structured query language (SQL)**

SQL is a programming language used to extract information from relational databases and to manipulate the tables in those databases (e.g., join them together and derive new tables).

Here are a couple of examples of the types of queries supported by SQL:

- List the position descriptions that have been applied for at Twitter.
- List the jobs that S. Walters has applied for.
- How many applications were received in the fourth quarter (Q4) of 2012?

From an analytical perspective, a key feature of SQL is that it can extract data from a relational database and put it into a flat tabular form more amenable to analysis.

An introductory statistics course can do no more than scratch the surface of this topic and show readers one source for data. The small example presented earlier is courtesy of Katya Vasilaky, who teaches a course on SQL and database queries at Statistics.com.

Big Data

Since the turn of the millennium, organizations have found that they have a lot of data already on their hands, or being continuously generated, that yield useful information simply by applying statistical and machine learning models:

- OKCupid, a dating site, uses statistical models with their data to predict what forms of message content are most likely to produce a response.
- Telenor, a Norwegian mobile phone service company, was able to reduce subscriber turnover by 37% by using models to predict which customers were most likely to leave and then lavishing attention on them.
- Allstate, an insurance company, tripled the accuracy of predicting injury liability in auto claims by incorporating more information about vehicle type.

The above examples are from Eric Siegel's *Predictive Analytics* (2013, Wiley).

In other cases, the flow of data can be harnessed for experiments that can be used as the basis for pricing decisions:

- Orbitz, a travel site, has found that it could price hotel options higher for Mac users than for Windows users.
- Staples online store found that it could charge more for staplers if a customer lived far from a Staples store.

The challenge of handling the data, though, is substantial. This challenge is not so much a function of the static size of the data, rather it results from the enormous flow of NEW data.

- Walmart, the large retailer, adds to its database more than 1,000,000 transactions per HOUR.
- JP Morgan reportedly made the decision in 2013 to retain financial data that it had previously been discarding after a set period; the result is that that it must add the equivalent of 2002 Terabyte disks DAILY.

Definition: **Big data**

Big data is a relative term—data are big by reference to the past and to the methods and devices available to deal with them. The challenge big data presents is often characterized by the four Vs—*volume*, *velocity*, *variety*, and *veracity*. *Volume* refers to the amount of data. *Velocity* refers to the flow rate—the speed at which it is being generated. *Variety* refers to the different types of data being generated (money, dates, numbers, text, etc.). *Veracity* refers to the fact that data are being generated by organic-distributed processes (e.g., millions of people signing up for services or free downloads) and not subject to the controls or quality checks that apply to data collected for a study.

For the practitioner of traditional statistical methods, big data introduce a whole level of complexity that was previously absent. A traditional large statistical research study might have involved, say, just 10–15 variables and 5000 records. The data would likely have been collected expressly for the purpose of conducting a study and the scarcity of data, or expense of obtaining it, would most likely have been the main issues.

If you consider the traditional statistical study to be the size of a period at the end of a sentence then the Walmart database is the size of a football field (Figure 1.2).

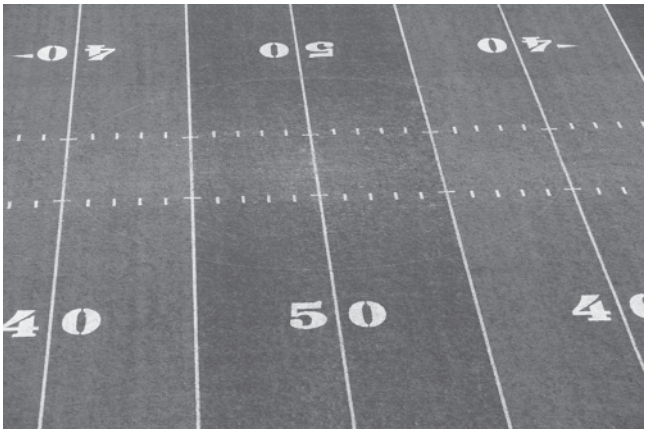


Figure 1.2 In comparing the Walmart database to a traditional statistical study, the difference in scale is like the difference in size between a football field (a real one, not the picture) and the period at the end of this sentence.

Implications for the Practice of Statistics and Statistics Professionals

For traditional research studies involving moderate amounts of data, little has changed about statistical practice or the jobs of statisticians. But the major job growth for statisticians since 2005 has been in the area of what is called *data analytics* or *data science*. Both are somewhat new terms and their definitions are hard to pin down. But central to both is the notion of using statistical and machine learning methods to extract useful information from available organizational data (often of huge size).

The great scale of the flow of new data means that the challenge of extracting, manipulating, cleaning, and preparing data is now enormous, and the time spent doing that easily outweighs the time spent analyzing data. The level of programming expertise required for these steps is substantial. Having gone to great lengths to prepare the data, adding some statistical algorithms into the process to gain interesting knowledge seems like a modest step to the programmer. As a result, statistical models are increasingly finding their way into the repertoire of computer scientists and IT professionals, and statisticians are increasingly called upon to apply their methods to big data.

1.10 VARIABLES AND THEIR FLAVORS

The third and fourth columns in Table 1.5 contain *variables*. These are things we observe, compute, or measure for each subject. They usually vary from one subject to the next. In

standard database format, each row represents an experimental unit or subject, while each column represents a variable. Two variables are missing from this table: the number of errors at the beginning of the study and the number at the end. From our point of view, these are intermediate steps.

Quantitative Variables

The numbers in the fourth column are the ones that really interest. They are an example of a *measurement variable* or *quantitative variable*. These are numbers with which you can do meaningful arithmetic. They fall into two types: discrete and continuous.

Definition: **Discrete variable**

The values in a discrete variable differ by fixed amounts and do not assume intermediate values.

The most common type of discrete variable is an integer variable, in which only integers are legal values. Family size is an example. More restricted discrete variables are often the result of rounding or choice of scale. For example, elevations might be any value, but their representations by contour lines on a topographic map are limited to intervals—for example, 50, 100, 150 foot contour lines.

Definition: **Continuous variable**

The values in a continuous variable can assume any values and the difference between any two values can be divided up into any number of legal values.

Age is a continuous variable as is elevation or longitude or latitude. Often, continuous variables may be binned into discrete variables for convenience, as with the contour lines on a map.

Categorical Variables

The other main type of variable is called *categorical* or *qualitative*.

Definition: **Categorical variable**

A categorical variable must take one of a set of defined non-numerical values—yes/no, low/medium/high, mammal/bird/reptile, and so on.

The binary data in column 3 (treatment or control) are categorical variables with two categories. Other examples that might be in the database (although not printed out earlier) are the city, county, or province of each hospital or whether it was a government, business, or charity hospital. Categorical data are often recorded in text labels; for example, male or female, Christian, Muslim, Hindu, Buddhist, Jew, or other. But it is also common to code categories numerically. In our database, treatment is a categorical variable and was coded as one. Control is coded as zero.

Do not do arithmetic on numerical codes for categorical data when it makes no sense. If we code Christian, Muslim, Hindu, Buddhist, Jew, or Other as 1, 2, 3, 4, 5, and 6, respectively, then finding the total or average of these codes is not likely to be meaningful. Coding qualitative data numerically does not make it quantitative! We will see later that some meaningful arithmetic *can* be done on categorical variables.

Computer code in programs will normally have to be told not only what type of variable to expect but also of any limitations on legal values (e.g., that age cannot be negative or that family size must be a positive integer).

Table Formats

Let us see how these data might be presented in other formats. One alternative is to present the error reduction for the control group in one column and the treatment group in the other. This provides the clearest presentation of how the two groups differ in the extent to which errors were reduced. The treatment group had, on average, 2.80 fewer errors in the second year. The control group had 1.88 fewer errors in the second year. Both groups had reduced errors, but the treatment group does appear to have reduced errors to a greater extent than the control group (Table 1.10).

TABLE 1.10 Error Reduction (Hypothetical)

Treatment	Control
2	3
2	1
5	2
2	2
2	1
2	1
4	1
2	3
2	1
2	4
4	1
2	2
3	1
9	5
2	1
2	4
2	1
2	1
3	2
2	2
6	2

TABLE 1.10 (*Continued*)

	Treatment	Control
	2	1
	2	1
	2	2
	2	2
Mean	2.80	1.88

The format in Table 1.10 is one that you might see in print; it might also be used by some softwares—especially Excel.

Technique

Excel's hypothesis testing procedures may require this format in which group is not a separate variable but instead is simply indicated by the column in which a value is located.

You may also see a format that is used only in print, in which all the data are displayed in successive rows for compactness. Neither the rows nor the columns have any significance, so this arrangement is not used with software (Table 1.11).

TABLE 1.11 The Hypothetical Hospital Error Reduction Data in Compact Format for Print

Treatment				
2	2	2	2	2
2	2	2	2	2
4	4	4	4	4
2	2	2	2	2
6	6	6	6	6
Control				
3	3	3	3	3
1	1	1	1	1
1	1	1	1	1
4	4	4	4	4
2	2	2	2	2

1.11 EXAMINING AND DISPLAYING THE DATA

Errors and Outliers are Not the Same Thing!

Suppose that Row 47 of Table 1.5 reads as follows:

Row	Hospital#	Treat?	Reduction in Errors
47	4076	10	2

10 is not a valid value for the "Treat?" variable, which needs to be either zero or one. We would then look up hospital number 4076 to see if we could find the reason for this error. Here is an example of a common error that might lead to a 10 in the Treat? column.

Imagine that we have a study with eight subjects and are supposed to type the eight values 1 0 0 1 0 0 1 0 into a list. Leaving out one space would give 1 0 0 10 0 1 0. As if that 10 were not bad enough, notice that the numbers for subjects six and seven are wrong now and there is no number for subject eight. Noticing this one outlier and trying to correct it helped us to find and fix three other errors.

Definition: **Outliers**

A value (for a given variable) that seems distant from or does not fit in with the other values for that variable is called an *outlier*. It could be an illegal value, as in this case. It could also be a very odd value or a legitimate one. If we saw a 456 in column four, this would not be illegal but it would be a very improbable degree of error reduction.

Some statistical software will identify outliers for you, but keep in mind that these are arbitrary identifications determined by arithmetic. Outliers are not necessarily errors—some are legitimate values. Consider these annual enrollments at a randomly selected set of 10 courses at Statistics.com.

8, 12, 21, 17, 6, 13, 29, 180, 11, 13

The 180 is certainly an outlier, but it is not incorrect. It is the enrollment in an introductory course, whereas the other enrollment figures are for more advanced courses.

Try It Yourself 1.12

Find the average enrollment for the 10 Statistics.com courses whose enrollments are listed above. Would you say this is a good representation of the typical enrollment?

Whenever we find an outlier, we need to investigate it and try to understand the reason for it. If there is an error, we need to try to correct it. Outliers, whether erroneous or legitimate, can strongly affect the numbers we compute from our data. In some cases, an outlier is a symptom of a deeper problem that could have an even greater impact on our results.

Outliers and Social Security

The Social Security Administration is a key source for wage data—Social Security taxes are due on almost all wages and employers must file earnings reports with the tax authorities.

One statistic reported regularly is the average pay of those receiving more than $50 million in wages. This number receives a great deal of attention in the policy debate over

income distribution. There were just 74 of these super-earners in 2009 and the government reported on October 15, 2010 that the average income of the super-earners more than quintupled in 2009—to an average of $519 million. This was quite an impressive feat during a severe recession, and the report came during a highly charged political atmosphere in which an important bone of contention was the relative share of income and wealth held by the richest members of society.

Shortly after the report was issued, analysts found that two individuals were responsible for this entire increase. Between them, these two taxpayers reported more than $32 billion in income on multiple W2 (tax) filings. As of November 3, 2010, no information was available on who the individuals were or why they reported such astronomical sums.

However, the Social Security Administration did determine that the filings from the two individuals were in error and issued a revised report. The results?

- 2009 super-earner average wages actually declined by 7.7% from 2008 instead of quintupling.
- 2009 average wages for all workers declined by $598 from 2008; the original report was $384.

These two outliers had a huge and misleading impact on key government statistics. They contributed $214 to the average income of all wage earners, and when they were removed, the recession's wage hit grew by more than 50%. At the same time, the fuel they added to the income distribution debate was illusory.

Question 1.5

What impact would these two outliers have had on statistics that used the median, rather than the mean?

Frequency Tables

Now that we have our results on 50 individual hospitals, we need a way to summarize and compare the treatments and controls as groups. We will look first at summaries that are numbers and then at summaries that are pictures. One numerical summary we could make here is a table of values and how often those values occur, that is, their frequencies (Table 1.12).

This is called a *frequency table* or frequency distribution.

***Definition:* Frequency table**

A frequency table is a table of possible values and the frequencies with which they occur in the data.

Let us interpret a couple of rows.

The bottom row tells us that there were 25 control group hospitals and 25 treatment group hospitals, for a total of 50 hospitals.

The first row tells us that 12 of the 25 control hospitals had a reduction in errors of one and that none of the treatment group hospitals had a reduction in errors of one.

TABLE 1.12 Frequency Distribution—Reduction in Errors

Value	Control	Treatment	Total
1	12	0	12
2	8	18	26
3	2	2	4
4	2	2	4
5	1	1	2
6	0	1	1
9	0	1	1
All	25	25	50

The second row tells us that eight of the 25 control hospitals had a reduction in errors of two and that 18 of the treatment group hospitals had a reduction in errors of two, so that a total of 26 hospitals had a reduction in errors of two.

Notice that if we did not have a control group, we would overestimate the success of the treatment. All the hospitals improved. Still, there is some good news. It looks like the treatment group showed more improvement.

Frequency tables often include *cumulative or relative frequencies*. Here is an example for the control group only (Table 1.13).

TABLE 1.13 Error Reduction Frequency Table (Control)

Error Reduction	Freq.	Cum. Freq.	Rel. Freq.
1	12	12	0.48
2	8	20	0.32
3	2	22	0.08
4	2	24	0.08
5	1	25	0.04
6	0	25	0.00
9	0	25	0.00
All	25	25	1.00

Here is row 2.

2 8 20 0.32

It illustrates that the value of two for reduction in medical errors showed up eight times in the control group, while a reduction of two *or fewer* errors showed up 20 times and the eight times out of 25 constituted 0.32=32% of the total.

Although it is not shown earlier, we could also calculate *cumulative relative frequency* in the same way that cumulative frequency is calculated—by adding together the current row with the preceding rows. For example, the cumulative relative frequency for the third row (error reduction = 3) is $0.48 + 0.32 + 0.08 = 0.88$.

Try It Yourself 1.13

Compute cumulative and relative frequencies for the treatment hospitals. Can you also find cumulative relative frequencies?

Histograms

We have been looking mostly at single values that summarize the data. Let us return to looking at all the data. Recall the frequency distribution presented earlier and turn it into a picture—a frequency histogram. Figure 1.3 shows a histogram of the error reductions for the treatment group.

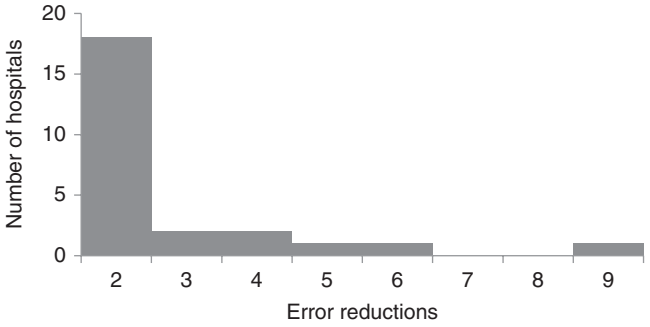


Figure 1.3 Frequency histogram of error reductions in the treatment group.

Interpreting Figure 1.3, we see that 18 hospitals reduced the number of errors by two, two hospitals reduced the number of errors by three, and so on. No hospitals reduced the number of errors by as much as seven or eight, but the histogram *must leave room* for these values to present an accurate picture.

In Figure 1.3, the histogram is relatively easy to make—there are only eight possible values, so we can have a vertical bar for each value.

If we are graphing more complex data—say, hospital sizes—we will not have enough room or visibility to devote one bar to each value. Instead, we group the data into bins. It is important that the bins be (i) equally sized and (ii) *contiguous*. By *contiguous*, we mean that the data range is divided up into equally sized bins, even if some bins have no data, like 7 and 8 as mentioned earlier. Consider Figure 1.4, which shows hypothetical data for hospital sizes in a mid-sized state.

We can see that there were 13 hospitals, with 0–99 beds, 14 hospitals with 100–199 beds, and so on. In Excel's native histogram function, the final bin on the right may include all values larger than a certain amount, but this has the disadvantage of not giving an accurate picture of the gaps in the data.

Deciding on how to display these data is not a trivial matter for a computer. The program must decide on how many values to place in a bin and where the bin boundaries are and the various forms of messiness that can arise. Often, the algorithm that is used results in non-integer values on the x-axis, which may not make sense. For example, the binning algorithm in one program produced the following histogram for the hospital error data (Figure 1.5). You can see that it makes much less sense than Figure 1.4.

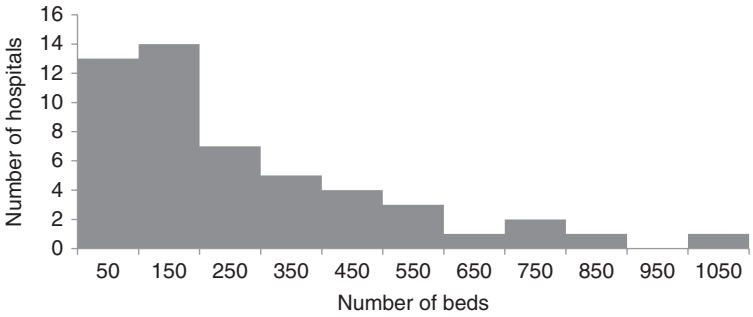


Figure 1.4 Hospital sizes by number of beds (hypothetical data for a mid-sized state).

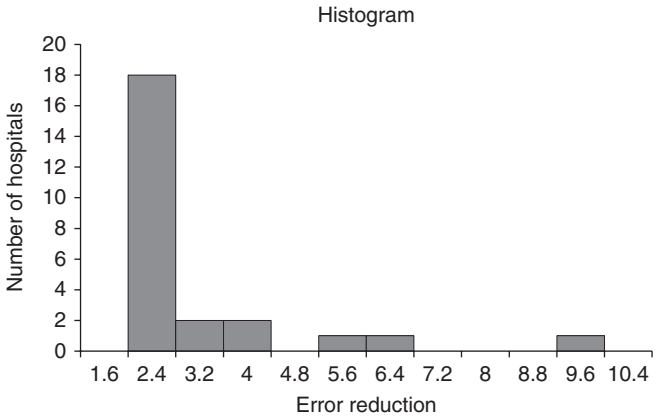


Figure 1.5 Histogram of hospital error reduction, treatment group (x values are not integers).

Figure 1.6 shows a *back-to-back* histogram of error reductions. It is described this way because it shows two sets of data—the control set and the treatment set—on the same line, with counts for the control set on the left side of the Errors column and counts for the treatment set on the right side. You may think of it as a horizontal histogram.

Stem and Leaf Plots

A variant of the histogram is the *stem-and-leaf* plot, in which the counts of x that you have seen earlier are replaced with numbers denoting the actual values. Figure 1.7 is a stem-and-leaf plot for hospital sizes in terms of number of beds in a hypothetical set of rural counties.

The column to the left of the vertical line indicates the "stem digit," which in this case represents units of 10. The numbers to the right indicate the "leaves," or the unit values. The number of digits on the right—two digits in row 1—tells us that we are counting two hospitals. In row 6, we see five digits, which means five hospitals.

Now, we read across to count the number of beds at each hospital. The first row tells us that there were two hospitals of size 23 [2|33]. The second row tells us there was one 35-bed hospital *and* one 39-bed hospital [3|59]. Again, row 6 tells us there are five hospitals, one with 74 beds, one with 76 beds, and three with 75 beds [7|46555]. The stem-and-leaf plot

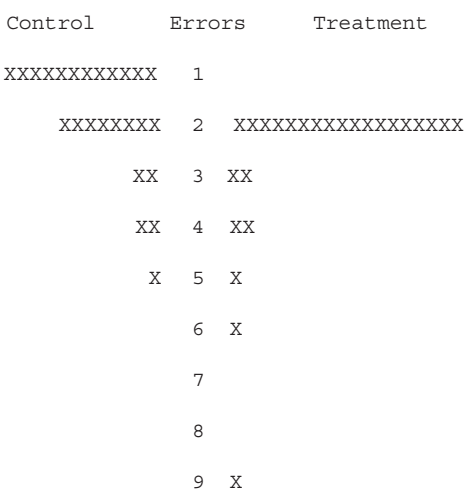


Figure 1.6 Back-to-back histogram.

2	33
3	59
4	69
5	555
6	179
7	46,555
8	23
9	5

Figure 1.7 Stem-and-leaf plot, hypothetical rural hospital sizes.

has the advantage of conveying more information than the histogram, but it is somewhat difficult to manage when the range in the number of digits in the data is larger than 2 or 3. Most software packages do not implement it easily.

Box Plots

Let us look at another graph of the data distribution—one that has features showing the percentiles of the distribution, as well as outliers.

With a boxplot,

- A central box encloses the central half of the data—the top of the box is at the 75th percentile and the bottom of the box is at the 25th percentile.
- The median is marked with a line.
- "Whiskers" extend out from the box in either direction to enclose the rest of the data or most of it. The whiskers stop when they come to the end of the data or when they get further than 1.5 *inter-quartile range* (IQR), from the top and bottom of the box—whichever comes first.
- Outliers beyond the whiskers are indicated with individual markers.

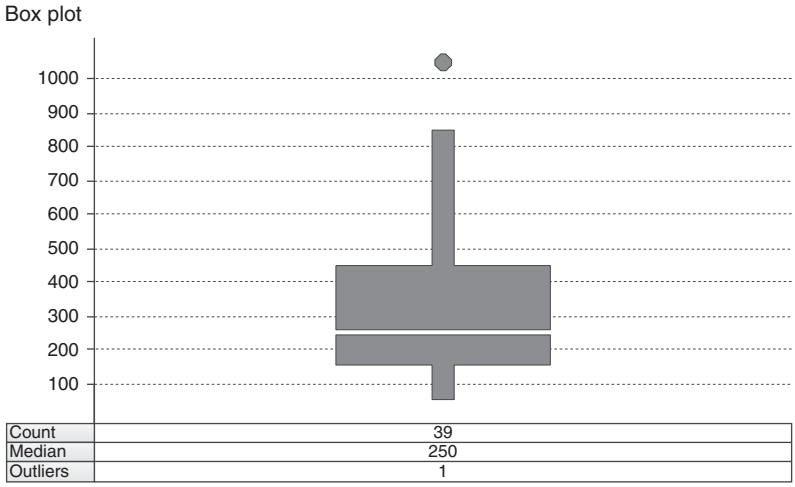


Figure 1.8 Boxplot of metropolitan area hospital sizes (y-axis shows the number of beds) (created using Spotfire).

Consider Figure 1.8, which is a boxplot of hospital sizes by number of beds in a hypothetical metropolitan area. We can glean the following information.

- Half the hospitals are between 150 beds and 450 beds.
- The IQR is 300 beds.
- The median hospital size is 250 beds.
- The rest of the hospitals are spread out between 50 beds and 850 beds, with the exception of one outlier hospital that has 1050 beds.

Boxplots are a compact way to compare distributions. Below is a side-by-side boxplot comparison of the reduction in errors for the control and treatment hospitals. It was created with XLMiner. Different software has varying ways of creating boxplots. In this case, XLMiner places lines on top of the whiskers, uses a + to indicate the mean, places horizontal V-shaped notches around the mean, and uses o to indicate an outlier (Figure 1.9).

Note how these boxes communicate information by the features that are missing—the top half of the box and the absence of lower whiskers.

Try It Yourself 1.14

What does the absence of the top half of the box and the lower whiskers communicate?

Tails and Skew

Let us review the picture we get from the histograms and the boxplots.

The location of the data is lower for the control group than for the treatment group, which is reflected in the value of the mean.

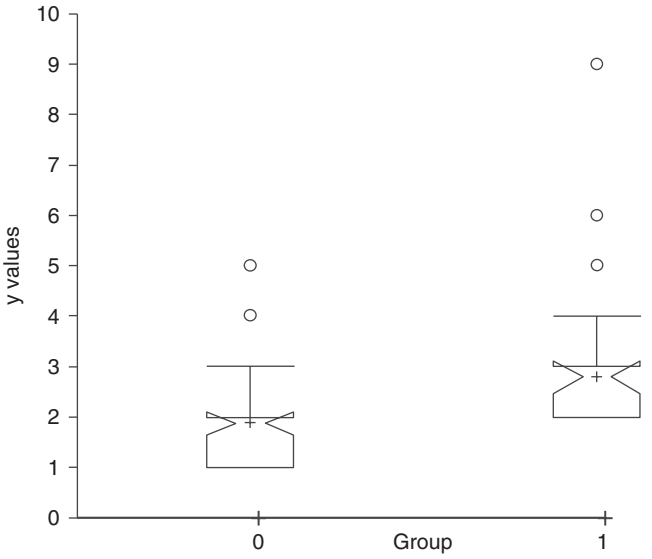


Figure 1.9 Error reductions (y-axis) for control hospitals (0) and treatment hospitals (1).

Other than the value of nine, the shape of the distribution for the treatment groups looks roughly like that for the control group. Both groups have peaks at the low end around one or two and trail off toward higher values. We call such a pattern *skewed toward high values*. The part of the picture where the data trail off, say around five, six, seven, eight, or nine, is called the *tail of the distribution*. The direction of the *skew* is the direction of the longer tail. The shape of the distribution is easier to see in the histogram than in the table.

Please look at the spreadsheet ErrorReductions.xls that summarizes all the relevant measures for both the treatment and control data. Be sure that you review and understand the formulas in the highlighted cells.

1.12 ARE WE SURE WE MADE A DIFFERENCE?

We found that the average effect of our treatment was to reduce the number of hospital errors by almost one—0.92. However, we see that the variability from hospital to hospital is more than one. Some people define statistics as the art and science of finding patterns in the midst of variability. We think we found a difference but there is enough variability that it is hard to be sure. For the original study reported by the CBC, we found that those results could well be due to chance. In the next chapter, we will try to determine if the difference we see in this current example is real or might simply be the result of random variability in the numbers.

APPENDIX: HISTORICAL NOTE

Before the widespread availability of computers and software, analysts relied on published tables of random numbers. Here is part of one such table (Table 1.14); the digits are arranged in separate groups of 5 for better visibility.

TABLE 1.14 Portion of a Random Digit Table

58535	99062	55182	89858	67701	94838	37317	10432	75653	78551
56329	09024	81507	90137	19241	55198	74006	52851	41477	58940
04016	38081	45519	27559	92403	30967	86797	17004	22782	09508
37331	94994	67305	34040	91360	83009	36925	31844	12940	51503
24822	53594	72930	23342	88646					

How can we use this table to do the work of 10 coin tosses? We need to convert each random digit into "heads" or "tails." There are several ways to do this. Here are two methods.

Odd = heads, Even = tails.

0–4 = heads, 5–9 = tails.

Let us select a random spot in the table and read off 10 digits. For example, looking at the left center of the table, we see

45,519 27,559

Using 0–4 = heads, 5–9 = tails, this amounts to

HTTHT HTTTT

The result is three heads and seven tails.

1.13 EXERCISES

1. Here are 20 more trials of the exercise you did at the beginning of the readings in which you investigated the model of a random distribution of 10 hospital errors between 2 years, 2008 and 2009. Each row is a trial—10 coin flips. You will use the results to determine whether it is unusual for 10 hospital errors over 2 years to be split 7–3 (2008, 2009), just by chance.

```
Run#
1     HHHTTHTTHH
2     TTHHHTTTHH
3     TTHHTHHTTT
4     HTTHTHHTTT
5     HTHTHTHTTT
6     TTHTTTTHHH
7     HHHTHTHHHH
8     HHTHHHTTTH
9     HHTTTTHTTT
10    THHTTTHTTH
11    TTHHHTTHHT
12    TTTHHHHTHT
13    TTHTHTTTTT
14    THTHHHTTTT
15    THTHHHTTTT
16    HHTTHTHHHH
17    HTTHTHTHTH
18    THHTHHHHHT
19    THHHTTHTTT
20    THTHHHTHTH
```

Each "H" or "T" represents an error. Under our chance model, let us say that "H" means the error happened in 2008 and "T" means 2009. Each row represents one trial (i.e., an allocation of 10 errors).

a. For each of the twenty runs, count the number of times "H" (2008) occurred.

b. Then make a frequency table for your results.

c. What proportion of the runs gave seven or more 2008s?

d. Comment on whether the difference between 2008 and 2009 might have happened by chance.

2. There is controversy over the effectiveness of surgery in treating prostate cancer. Give a design for a study to address this issue. You should address all the issues raised in this week's lesson. For each issue either suggest a way to address that issue or give a reason why there may be no way.

3. Below is a list of numbers of home runs hit by the home runs leaders of the American league during the years 1951–1965:

```
33,52,43,32,37,52,42,42,42,40,61,48,45,49,32
```

Calculate mean, variance, and standard deviation of the data. Note: You can use standard Excel functions or any other software of your choice here. Although this is a small data set, most people would consider that it is the population of interest here—not a sample from some much larger population.

4. Estimate the probability that a family with 10 children would have three or fewer girls. Explain how you arrived at your estimate. Hint: Assume that each successive child has a 50/50 probability of being a male or female, and use dice, coins, or random numbers, not a theoretical formula.

5. Consider the following three customers and the items they have purchased from an online merchant in the last month. A "1" indicates that item was purchased, a "0" indicates that item was not purchased.

	Book	MP3	Power Tool	Tablet	Game
Cust 1	1	0	0	1	1
Cust 2	1	0	1	0	0
Cust 3	0	0	0	1	1

a. Calculate all the possible Euclidean distances between customer pairs.

b. Suppose that you now want to set up a recommender system that alerts a customer when a new item is purchased by a like-minded customer. Which customer would be a good source of recommendations for customer 1?

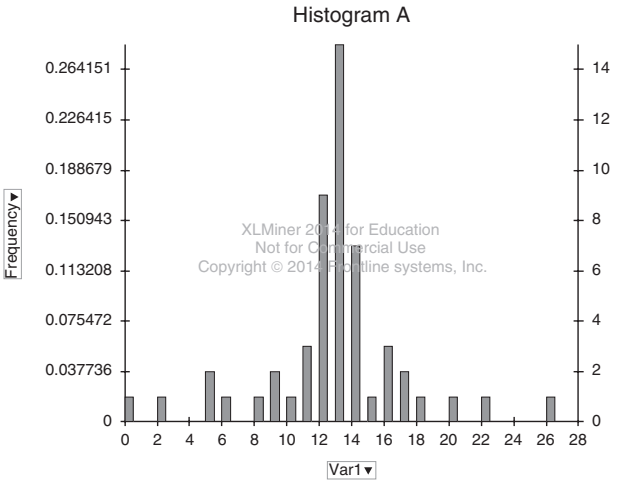
c. Imagine now that you have millions of items available for purchase, and tens of millions of customers. How could you use Euclidean distance as part of a system to keep a customer from being inundated with recommendations?

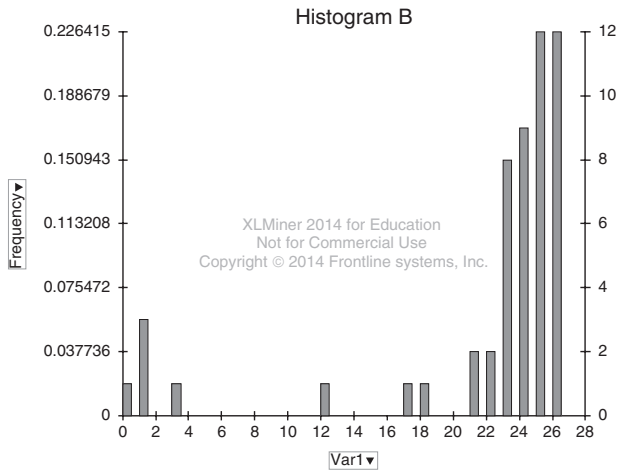
Use the PulseNew.xls data for all the remaining questions (see the book website for these and all data sets):

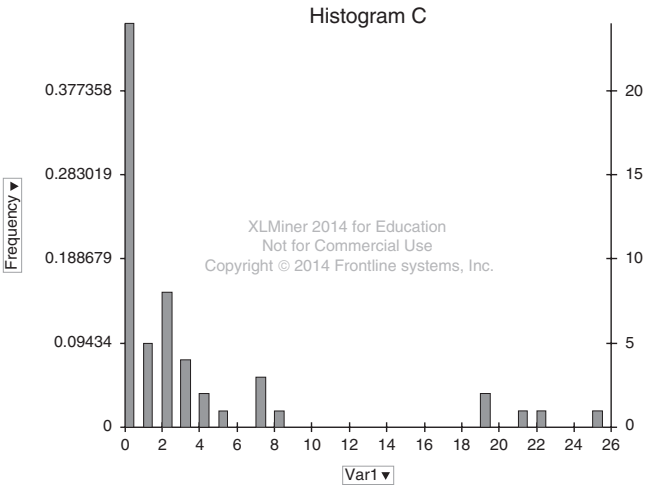
6. Make a frequency table for the variable Ran? Describe what you find briefly in words. Although this can be done by hand, this is a good opportunity to get acquainted with your software program. HINT: In Excel, try the COUNTIF function (make sure Analysis Toolpak and Analysis Toolpak VBA are installed). In Statcrunch, use Stat > Tables > Frequency.

7. The numbers for how many ran and how many did not run seem a bit out of balance. Explain how you would check to see if this could reasonably be considered only due to chance. (Just explain the plan; you do not have to carry it out.) Hint: Rolling dice, tossing coins, or using a random number table might be part of your strategy.

8. Consider a metropolitan area with a diversity of neighborhoods (commercial, shopping, residential, industrial, etc.). Restaurants are not spread evenly throughout the area—they tend to be located in a small number of "restaurant districts." Below are three possible histograms for the distribution of restaurants per neighborhood (number of restaurants is on the *x*-axis). Which best represents what the above-described distribution of restaurants looks like?







Answers to Try It Yourself

1.1 Results will vary, but chances are this happened at least once.

1.2 Results will vary; in a large class about half the students will get a trial with 14 or more heads, the rest will not.

1.3 You would need to get the subjects to agree to watch TV according to your orders. This might be feasible for a study of short-term effects, but it would be difficult to do if the study required a very long-term period (months or years). Also, some subjects (or their parents) might be averse to watching violent TV, causing dropouts.

1.4 Certainly, the subjects will know what TV shows they are watching. You may be able to conceal that you are studying violence. Those who follow the subjects to measure violent behavior in the future probably can be blind to which subjects received which treatments.

1.5 So many different factors might affect violent behavior that it would be difficult to even list them all, let alone control them.

1.6 A control group could watch no television or watch only nonviolent programs. If we have data for the subjects on other variables we think might also impact violent behavior, then we could pair subjects who are similar on those other variables

1.7 The range is $9 - 2 = 7$.

1.8 Control group IQR: $2 - 1 = 1$
Treatment group IQR: $3 - 2 = 1$
IQR for all observations: $2 - 2 = 0$

1.9

x	mean	Residual	Residual^2
8	4	4	16
1	4	-3	9
4	4	0	0
2	4	-2	4
5	4	1	1

The mean is $(8+1+4+2+5)/5 = 20/5 = 4$. The variance is the average of the sum of the squared residuals = 6. The standard deviation is the square root of the variance = 2.45. (Note that if this was a sample from a larger population, in calculating the variance, we would have divided by $n-1=4$.) The standard deviation is within the range of the residuals, so it is "in the ballpark."

1.10 If you did enough resamples, the mean of the resample variances will be smaller than the "population" variance.

1.11 Hours spent watching television should be relatively easy to measure but is that what you really want to measure? If we try to determine how much violence each subject watches, we have to find some way to define and measure that. There are a number of websites such as http://www.kids-in-mind.com/help/ratings.htm that offer some rating on a scale of 0–10, depending on quantity as well as context.

1.12 The average enrollment is 31. This is well above the enrollment for all but one course, so it is not typical. The average is skewed high by the enrollment in one large course—introductory statistics.

1.13 Cumulative, relative, and cumulative relative frequencies

Value	Frequency	Cum. Freq.	Rel. Freq.	Cum. Rel. Freq.
2	18	18	0.72	0.72
3	2	20	0.08	0.8
4	2	22	0.08	0.88
5	1	23	0.04	0.92
6	1	24	0.04	0.96
9	1	25	0.04	1.00
All	25		1.00	

1.14 They indicate that the bulk of the data is "bunched up" at 1–2 (for the control group) and 2–3 for the treatment group.

Answers to Questions

1.1 Probabilities must always lie between 0 and 1.

1.2 This would be equivalent to 10 heads in 10 tosses, very unlikely, so chance is not a reasonable explanation.

1.3 The student failed to sort the data before picking the middle value. If the sorting had been done, the numbers would be arranged 3, 4, 5, 7, 9 so you would see that the median is 5.

1.5 The median would not be greatly affected by these outliers as their presence or absence would hardly affect the value of the middle observation of dozens of observations (for the super-earners) or millions of observations (for the entire wage-earning population).