Chapter 1

WHAT IS STATISTICS?

Commonly, the word *statistics* means the arranging of data into charts, tables, and graphs along with the computations of various descriptive numbers about the data. This is a part of statistics, called **descriptive statistics**, but it is not the most important part. The most important part is concerned with reasoning in an environment where one does not know, or cannot know, all of the facts needed to reach conclusions with complete certainty. One deals with judgments and decisions in situations of incomplete information. In this chapter, we will give an overview of statistics along with an outline of the various topics in this book.

1.1 SAMPLING AND ESTIMATION

Let's begin with an example.

Example 1.1 ***Harris Poll.*** Louis Harris and Associates[1] conducts polls on various topics, either face-to-face, by telephone, or by the Internet. In one survey

[1] www.harrisinteractive.com

on health trends of adult Americans conducted in 1991, surveyors contacted 1256 randomly selected adults by phone and asked them questions about diet, stress management, seat belt use, and so on. One of the questions asked was "Do you try hard to avoid too much fat in your diet?" Harris reported that 57% of the people responded Yes to this question. The article stated that the margin of error of the study was plus or minus 3%.

This is an example of an inference made from incomplete information. The group under study in this survey is the collection of adult Americans, which consists of more than 200 million people. This is called the **population**. If every individual of this group were to be queried, the survey would be called a **census**. Yet of the millions in the population, the Harris survey examined only 1256 people. Such a subset of the population is called a **sample**.

Once every ten years, the U.S. Bureau of the Census conducts a survey of the entire U.S. population. The year 2000 census cost the government billions of dollars. For the purposes of following health trends, it's not practical to conduct a census. It would be too expensive, too time-consuming, and too intrusive into people's lives. We shall see as we progress through this text that, if done carefully, 1256 people are sufficient to make reasonable estimates of the opinion of all adult Americans. Samuel Johnson was aware that there is useful information in a sample. He said that you don't have to eat the whole ox to know that the meat is tough.

The people or things in a population are called **units**. If the units are people, they are sometimes called **subjects**. A characteristic of a unit (such as a person's weight, eye color, or the response to a Harris Poll question) is called a **variable**. If a variable has only two possible values (such as a response to a Yes or No question, or a person's sex) it is called a **dichotomous variable**. If a variable assigns a number to each individual (such as a person's age, family size, or weight), it is called a **quantitative variable**.

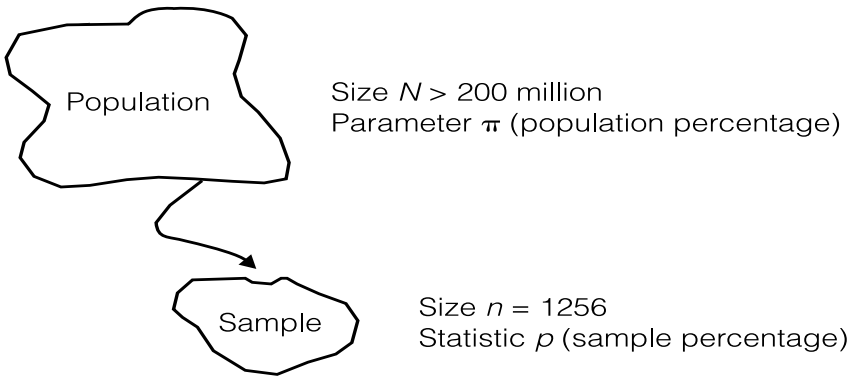

Figure 1.1 Parameter and statistic for a dichotomous variable

A number derived from a *sample* is called a **statistic**, whereas a number derived from the *population* is called a **parameter**. Parameters are usually denoted by Greek letters, such as π, for population percentage of a dichotomous variable, or μ, for population mean of a quantitative variable. For the Harris study, the **sample percentage** $p = 57\%$ is a statistic. It is not the (unknown) **population percentage** π, which is the percentage that we would obtain if it were possible to ask the same question of the entire population.

Inferences we make about a population based on facts derived from a sample are uncertain. The statistic p is not the same as the parameter π. In fact, if the study had been repeated, even if it had been done at about the same time and in the same way, it most likely would have produced a different value of p, whereas π would still be the same. The Harris study acknowledges this variability by mentioning a margin of error of $\pm 3\%$.

How can the researchers say that the margin of error is plus or minus 3% when such a small sample of all adult Americans was contacted? This is one of the questions that we will deal with in this book.

1.2 SIMULATION

Consider a box containing chips or cards, each of which is numbered either 0 or 1. We want to take a random sample from this box in order to estimate the percentage of the cards that are numbered with a 1. The population in this case is the box of cards, which we will call the **population box**. The percentage of cards in the box that are numbered with a 1 is the parameter π. In the Harris study, the parameter π is unknown. Here, however, in order to see how samples behave, we will make our model with a known percentage of cards numbered with a 1—say, $\pi = 60\%$. At the same time, we will estimate π, pretending that we do not know its value, by examining 25 cards in the box.

We take a **simple random sample with replacement** of 25 cards from the box as follows. Mix the box of cards, choose one at random, record it, replace it, and then repeat the procedure until we have recorded the numbers on 25 cards. Although survey samples are not generally drawn *with replacement*, our simulation simplifies the analysis because the box remains unchanged between draws; so, after examining each card, the chance of drawing a card numbered 1 on the following draw is the same as it was for the previous draw—in this case a 60% chance. Let's say that, after drawing the 25 cards this way, we obtain the following results, recorded in 5 rows of 5 numbers:

0	1	1	1	1
1	0	1	1	0
1	0	1	0	1
0	0	0	0	1
1	0	1	0	1

Based on this random sample of 25 draws, we want to guess the percentage of 1's in the box. There are 14 cards numbered 1 in the sample. This gives us a sample percentage of $p = 14/25 = 0.56 = 56\%$. If this is all of the information we have about the population box, and we want to estimate the percentage of 1's in the box, our best guess would be 56%. Notice that this sample value $p = 56\%$ is 4 percentage points below the true population value $\pi = 60\%$. We say that the **random sampling error** (or simply **random error**) is -4%.

Equivalently, instead of using a box of cards, we can simulate this experiment by generating random numbers using a programmable calculator, such as the **TI-83 Plus**, a computer program, such as **Excel** or **Minitab**, or a table of random digits such as Table 1 on page **??**. In later chapters, we will give instructions for using technology in sections marked **Notes on Technology**, but for now we will use a table of random digits to demonstrate the procedure. For convenience, the random digits in Table 1 on page **??** are grouped into fives. Each of the 45 rows holds 10 groups of fives. The entire table holds $45 \times 10 \times 5 = 2250$ random digits. To guard against using the same random numbers in every simulation, we should start in a random row and column—say, row 16, column 6. We can obtain a random starting point by tossing a paper clip or coin on the page. The first 25 digits in row 16, starting with column 6, are

99109 14827 24949 16210 95105

Because we want a 60% chance of drawing a 1, we assign the value 1 to the six random digits $0, 1, 2, 3, 4, 5$ and assign the value 0 to the four digits $6, 7, 8, 9$; this results in the following values:

00110 11010 11010 10111 01111

In this simulation, we ended up with 16 1's, which results in the statistic $p = 16/25 = 0.64 = 64\%$. The random sampling error is $+4\%$.

Continuing where we left off in row 16, the next 25 digits go into row 17 and result in

00100 00101 01000 11111 11110

This time we have 13 1's to obtain a statistic of $p = 52\%$ and a random sampling error of -8%.

One more time:

11101 11110 11110 11101 01111

gives us 20 1's with $p = 80\%$, and a random error of $+20\%$.

The random errors were -4% from the box of cards and $+4\%, -8\%$, and $+20\%$ from the table of random numbers.

1.3 ERROR ANALYSIS

An **experiment** is a procedure that results in a measurement or observation. The Harris Poll is an experiment that resulted in the measurement (statistic) of 57%. An experiment whose outcome depends upon chance is called a **random experiment**. On repetition of such an experiment, one will typically obtain a different measurement or observation. So, if the Harris Poll were to be repeated, the new statistic would very likely differ slightly from 57%. Each repetition is called an **execution** or **trial** of the experiment.

The four simulations above are trials of a random experiment that resulted in four different percentages. The random sampling errors of the four simulations average out to

$$\textbf{AV:} \quad \frac{-4\% + 4\% - 8\% + 20\%}{4} = +3\%$$

Note that the cancellation of the positive and negative random errors results in a small average. Actually, with more trials, the average of the random sampling errors tends to zero.

So in order to measure a "typical size" of a random sampling error, we have to ignore the signs. We *could* just take the **mean of the absolute values (MA**) of the random sampling errors. For the four random sampling errors above, the MA turns out to be

$$\textbf{MA:} \quad \frac{|-4\%| + |+4\%| + |-8\%| + |+20\%|}{4} = 9\%$$

The MA is difficult to deal with theoretically because the absolute value function is not differentiable at 0. So in statistics, and error analysis in general, the **root mean square (RMS)** of the random sampling errors is generally used. For the four random sampling errors above, the RMS is

$$\textbf{RMS:} \quad \sqrt{\frac{(-4\%)^2 + (+4\%)^2 + (-8\%)^2 + (+20\%)^2}{4}} = \sqrt{124}\,\% = 11.14\%$$

The RMS is a more conservative measure of the typical size of the random sampling errors in the sense that MA $\leq$ RMS (see Review Exercise 1.21 at the end of this chapter).

For a given experiment, the RMS of *all* possible random sampling errors is called the **standard error (SE)** . We will study the standard error more formally later in the book. For example, whenever we use a random sample

of size n and its percentages p to estimate the population percentage π of a dichotomous population, we will show that

$$\mathrm{SE}_p = \sqrt{\frac{\pi(1-\pi)}{n}} \leq \frac{1}{2\sqrt{n}} \tag{1.1}$$

which for $n = 25$ comes out to $\mathrm{SE} = 0.097979\ldots \leq 0.10$. Notice that the estimate $\mathrm{SE}_p \leq 10\%$ is reasonably close to the RMS $= 11.14\%$ that we obtained in our four simulations. As the number of simulations increases, the RMS will tend to the SE. A proof of the inequality in Formula 1.1 is an exercise in calculus and is left to the reader as Review Exercise 1.13. This inequality is useful because, in a situation such as the Harris Poll, the value of the parameter π is generally unknown.

Notice from Formula 1.1 that, other things being equal, as the size n of the random sample goes up, the standard error goes down in proportion to the square root of the sample size. So for samples four times as large, the standard error will be $1/\sqrt{4} = 1/2$ as large. This means that, for samples of size 100, the standard error will be no more than approximately $\frac{1}{2} \times 10\% = 5\%$. If we combine our four samples in our simulations of size 25, we notice that we have a total of $14 + 16 + 13 + 20 = 63$ cards numbered 1, which gives us $p = 63\%$, and thus a random error of $+3\%$. Another sample of 100 would likely give us a different random sampling error.

Although the standard error will be studied in more detail in later chapters, at this point we want to observe an interesting property. The standard errors of independent trials of a random experiment have a Pythagorean relationship, just as do right triangles in geometry. If A is a measurement of a parameter α with a standard error SE_A, and B is an independent measurement of a parameter β with a standard error SE_B, then $A + B$ is a measurement of $\alpha + \beta$ with a standard error $\sqrt{{\mathrm{SE}_A}^2 + {\mathrm{SE}_B}^2}$. In error analysis, it is convenient to write

$$A = \alpha \pm \mathrm{SE}_A$$

$$B = \beta \pm \mathrm{SE}_B$$

and

$$A + B = \alpha + \beta \pm \sqrt{{\mathrm{SE}_A}^2 + {\mathrm{SE}_B}^2}$$

Note that the measurements add, but not the SE's. This property results in one of the important parts of the law of averages.

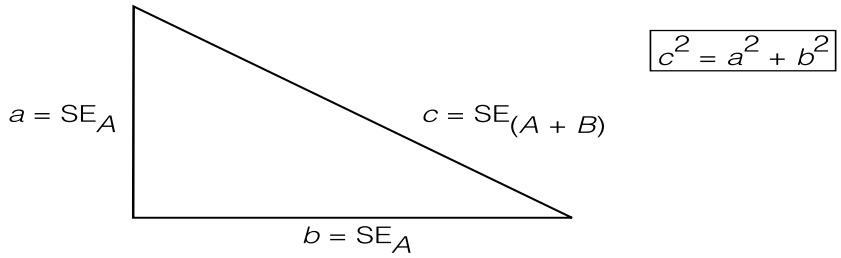

Figure 1.2 Pythagorean property of measurement error

To illustrate this relationship, let's say that a person commutes to work by car. She uses the car's trip odometer to measure the distance traveled. After many trips to work, she observes an average reading of 24.3 miles. Due to changes in traffic, weather, and other conditions, each trip results in a slightly different odometer reading, from which she estimates that the SE is approximately 0.4 miles. If A represents the distance measured on any given trip, we can write

$$A = 24.3 \pm 0.4 \text{ miles}$$

This does not mean that *every* trip varies by 0.4 miles; rather, it means that for such trips, odometer readings are approximately 24.3 miles and vary by about 0.4 miles, on the average.

She also often travels to a cottage on weekends. The distance to the cottage is greater than the commute, and the odometer readings are more variable because the trips involve stops for fuel and fast food. Let's say that an odometer reading B for a trip to the cottage has the formula

$$B = 67.2 \pm 1.1 \text{ miles}$$

A trip from work followed by a trip to the cottage will have a total odometer reading C, which can be calculated by adding the readings A and B. The Pythagorean relationship results in

$$\begin{aligned} A + B &= 24.3 + 67.2 \pm \sqrt{(0.4)^2 + (1.1)^2} \text{ miles} \\ C &= 91.5 \pm 1.2 \text{ miles} \end{aligned}$$

EXERCISES FOR SECTION 1.3

• **1.1** A medium apple has an average of 80 calories. However, the number of calories A varies with the size and type of apple. Let's say that the standard error

is $\mathrm{SE}_A = 10$ calories. Similarly, the number of calories B in a bran muffin is 400 calories with a standard error of $\mathrm{SE}_B = 40$ calories. Let C be the total number of calories in a lunch consisting of a medium apple and a bran muffin.

a. Find the estimated total number of calories of C along with the standard error SE_C.

b. Do the same for a lunch that also includes a cup of yogurt, which has 220 calories, give or take 15 calories or so.

1.2 For a biology lab experiment, you have to randomly select four mice. The mice to be chosen come from a population bred to weigh 30 grams, give or take 5 grams or so.

The total weight T of the four mice will be _____ grams, give or take _____ grams or so. Use the Pythagorean property of standard errors.

1.3 Repeat the simulation described in this section by taking another four samples of size $n = 25$ from a population box with $\pi = 0.60$. Use either Table 1 of random numbers[2] or a calculator with a RAND function.[3]

a. What average value of p do you get?

b. What average value of the random sampling errors do you get?

c. What MA of the random sampling errors do you get?

d. What RMS of the random sampling errors do you get?

C **1.4** Use a computer program or a programmable calculator to simulate 100 samples of size $n = 25$ from the above population box with $\pi = 0.60$. The values of p will vary from simulation to simulation.

a. What average value of p do you get?

b. What average value of the random sampling errors do you get?

c. What MA of the random sampling errors do you get?

d. What RMS of the random sampling errors do you get?

Students doing this experiment will have varying results, but the RMS should be close to that predicted by Formula 1.1 on page 6 above.

C **1.5** Continuing with Exercise 1.4, do the same for samples of size $n = 100$.

C **1.6** Continuing with Exercise 1.4, do the same for samples of size $n = 1256$.

[2]Obtain a starting point by tossing a coin or paper clip on the table on page **??**.

[3]The RAND function on a graphing calculator has to be initialized on its first use. You can use your Social Security number as a seed. Initializing is like choosing a random starting point on the table. Do this only once, otherwise the calculator will produce the same sequence of numbers every time.

1.4 PRECISION VERSUS CONFIDENCE

The "margin of error" of 3% mentioned in the Harris survey article is not the standard error. It should be more precisely stated as the margin of error for 95% confidence; that is, approximately 95% of the time, the random error will be within this "margin of error." In the news media, the 95% condition is generally assumed but not always explicitly mentioned. We will show in later chapters, that, for large random samples as in the Harris survey, this 95% margin of error is approximately 1.96 times (roughly *twice*) the standard error. Using the inequality in Formula 1.1 on page 6, we note that for $n = 1256$, we obtain $\text{SE}_p \leq \frac{1}{2\sqrt{1256}} = 0.0141 = 1.41\%$, which is close to half of the researchers' reported 3%. Because they obtained a statistic of $p = 57\%$, they state that they are 95% confident that $p = 57\%$ is within 3 percentage points of the true, but unknown, parameter π.

The margin of error is a measure of the **precision** of the estimate. However, it is inversely related to the level of **confidence** in this estimate. For a given sample size, if we want to *increase* confidence, we have to *decrease* the precision, and vice versa. For example, we can be 100% confident that π is somewhere between 0% and 100%. Although our confidence is high, the precision is so poor, it makes the statement worthless. On the other hand, we will show later that a margin of error of $\pm$*three* standard errors, which gives us a precision of $\pm 3\frac{1}{2\sqrt{1256}} = \pm 0.0423 = \pm 4.23\%$, will give us a statement with 99.7% confidence. And to get a statement with 99% confidence we need a margin of error of approximately ± 2.6 standard errors, which is approximately $\pm 3.6\%$. Using a margin of error of 1 standard error will result in a statement with only 68% confidence. At this point, do not worry about the specific percentages and levels of confidence. Details and derivations will come in later chapters.

In statistics, for a given data set, precision is inversely related to the level of confidence. However, by increasing the sample size both precision and level of confidence can be improved.

This Harris Poll example shows how an estimate of an unknown parameter is made by examining a random sample. This is part of the study of **inferential statistics**. Our simulation, on the other hand, involves a population box whose content is known. Examining the different samples that are possible and likely from a population with a known parameter is part of **sampling theory**. An understanding of the behavior of samples from a known population is a prerequisite of inferential statistics.

Note on estimation errors

It is important to realize that random sampling error is not the only source of error in a survey.[4] In Chapter **??** on page **??**, we comment on a news brief by Susan Hill, Senior Program Analyst at the National Science Foundation Division of Science Resource Studies. This news brief includes data on the number and types of degrees offered by postsecondary schools in the United States. The target population consists of all 2-year and 4-year institutions of higher learning in the United States. This survey is actually a census, because all institutions accredited to award degrees are queried in the initial mailing (close to 6000 in all). Follow-ups are conducted by means of mail and telephone. One source of error is **nonresponse**. Nonresponse rates for institutions ranged from 4% to 15%. The study attempted to correct for this by the imputation of missing values using prior-year returns, whenever these were available.

This report also discusses nonresponse rates for racial and ethnic data when bachelor's degree recipients were queried. The report estimates that for the year 1991, racial and ethnic nonresponse rates ranged from a low of 2.4% of the white, non-Hispanic recipients to a high of 14.9% of the Asian or Pacific Islander recipients. Nonresponse is a problem encountered whenever a survey or census is attempted. As you see from this example, nonresponse rates are not uniform across all segments of a population, so that nonresponse could affect some groups more than others, **biasing** the results.

Measurement error can also bias a survey. In this study, degrees are grouped in various ways (science and engineering degrees, chemistry, computer science, etc.). Institutions in different regions may use groupings of degrees (**taxonomies**) that differ from those of the survey. They may classify double majors or dual B.S. and M.S. degrees differently. The institutional representatives who fill out the survey forms may find the survey's taxonomy difficult and confusing. This could cause measurement errors.

Another source of bias is **response bias**, which occurs when a respondent gives an incorrect response. The respondent may be influenced by the phrasing of the question, or may not recall something correctly, or may simply be lying.

Whereas random sampling error can be reduced by increasing the size of the sample, sampling bias cannot. Sampling bias can only be reduced by changing the method of collecting the data.

[4] For example, the National Science Foundation publication NSF 95-318, *Guide to NSF Science and Engineering Resources Data*, describes several ongoing surveys designed to obtain information about science education and funding in the United States. This report includes a frank discussion of potential errors in these studies.

EXERCISES FOR SECTION 1.4

• **1.7** As part of a project for a political science class, Tina and Karen plan to take a random phone survey similar to the Harris Poll, except that they want to estimate the public sentiment on a bill before the U.S. Congress.

a. Use Formula 1.1 on page 6 to estimate the size needed for their sample if they want results with a standard error of less than 5%.

b. What size should the sample be if they want the 95% margin of error to be less than 5%?

c. What size for the 99.7% margin of error to be less than 5%?

1.8 In another Harris Poll of 1144 adult Americans, 306 people felt that the U.S. Constitution should be amended to have presidential elections decided by popular vote rather than by the electoral college. Find the statistic p and estimate the standard error SE_p.

• **1.9** As part of a 1990 study on the causes of asthma, the parents of 939 seven-year-old children in five German cities were interviewed.[5] Of these children, 57 had had a doctor's diagnosis of asthma at some time in their lives. Find the statistic p and estimate the standard error of this statistic.

1.10 *The New York Times* reported on a poll conducted October 18-21, 2000. This random phone survey found that among 1010 registered voters, 45% responded Yes to the question "Does the presidential candidate George W. Bush have the ability to deal wisely with an international crisis?" *The New York Times* explained, "In theory, in 19 cases out of 20 the results based on such a sample will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all American adults." Fill in the following, choosing from these options:

population, parameter, sample, statistic, unknown, π*,* p*,* 45%*,* 3%*,* 1.5%

The symbol for the parameter is _____. The value of the parameter is _____. The collection of 1010 registered voters is called the _____. The symbol for the statistic is _____. The value of the statistic is _____. The standard error is approximately _____. The 95% margin of error is approximately _____.

[5] Susanne Lau et al., *Early exposure to house-dust mite and cat allergens and development of childhood asthma: A cohort study*, The Lancet **356** (2000), pp. 1392–97.

1.5 HOW HYPOTHESES ARE TESTED

Let's consider an example along with two hypotheses.

Example 1.2 ***Salk vaccine trials.*** According to an article by Paul Meier[6], over a million children participated in a trial in 1954 of the Salk vaccine to see whether it would protect children against polio. In one part of the study, as summarized in Table 1.1 below, 401,974 children were injected with either the Salk vaccine or a salt solution placebo. The injections of the vaccine and placebo were assigned to the children at random. Furthermore, the trial was **double-blind**; that is, neither the children nor the diagnosing physicians were aware of who had been given vaccine or the salt solution.

Table 1.1 Summary of randomized double-blind Salk vaccine trials

Group	Number of subjects	Fatal polio	Paralytic & fatal polio	Nonparalytic polio	Total cases
Vaccine	200,745	0	33	24	57
Placebo	201,229	4	115	27	142
Totals	401,974	4	148	51	199

Does Salk vaccine prevent death from polio?

None of the children who received the Salk vaccine died of polio whereas four of the children in the placebo group died of polio. This seems to be evidence for the effectiveness of the vaccine, but how strong is the evidence? Polio was not a common disease, and the incidence of it would vary from year to year. Before the vaccine was developed, it was conceivable to have only four deaths in a group of approximately 400,000, which is a rate of 1 per 100,000.

For the sake of argument, let us suppose that the Salk vaccine was completely ineffective in preventing death from polio (this is called the **null hypothesis**); that is, suppose the vaccine prevented no deaths from polio, so that the four unfortunate children who died of polio just happened to fall into the placebo group by chance.

Assuming the null hypothesis, the vaccine had no effect and the four children would have died with or without the vaccine. The four children fell into the placebo group by chance, as if by the toss of a fair coin—say, heads for vaccine and tails for placebo. What is the chance that, for all four of the children who died, the coin came up tails? It is certainly not impossible for a coin to come up tails four times in a sequence. The chance of tails on each toss of a fair coin is 50%, so the chance of four tails in sequence is 50% of

[6]Paul Meier, *The Biggest Public Health Experiment Ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine.* This article appears in Judith Tanur, et al. (editors), *Statistics a Guide to the Unknown*, third edition, Duxbury Press, 1989.

50% of 50% of 50%, or $(0.50)^4 = 0.0625 = 6.25\%$. Such a low percentage is evidence for the effectiveness of the Salk vaccine in preventing death, but the evidence is not overwhelming. It would *not* be considered to be proof beyond a reasonable doubt.

This does not show that the vaccine is ineffective in preventing death from polio. It just means that there is insufficient evidence to conclude that the vaccine prevents death from polio.

> ***Absence of evidence is not evidence of absence.***[7]

Does it prevent paralytic polio?

Of the children who were injected with the Salk vaccine, 33 were later diagnosed with paralytic polio compared with 115 of the children in the placebo group. Using a similar analysis, we presume the null hypothesis that the vaccine was completely ineffective against paralytic polio; that is, the $33+115 = 148$ would have gotten paralytic polio with or without the vaccine. This is as if 148 tosses of a fair coin come up 33 heads and 115 tails. In 148 tosses, you expect approximately half of them to be heads; that is, 74 should be heads or so. But what is the chance that 148 tosses result in as few as 33 heads? Although it is not impossible to get so few heads, we shall see later when we study **probability theory** that the probability is less than 1 in 255 billion. Given this evidence, retaining the null hypothesis would be outrageous. We have no reasonable choice but to reject the null hypothesis and conclude that the Salk vaccine is effective against paralytic polio.

1.6 PREVIEW

Here are some of the issues that we will consider in this book:

- How do you take a sample? What is a random sample? In the Harris study, how do you account for people who cannot be reached by phone, do not answer the phone, refuse to participate, misunderstand the question, lie, and so on? These are all parts of the topic of the **design of experiments**. Although no chapter of this text is dedicated exclusively to this topic, the issues involved are pointed out throughout the book.

[7] Douglas G. Altman, J. Martin Bland, *Statistical Notes: Absence of evidence is not evidence of absence*, British Journal of Medicine **311** (1995), p. 485. This article can be found on the Web at http://bmj.com/cgi/content/full/311/7003/485.

- Once you have a sample, which consists of a collection of data, you want to organize and summarize this data. This can be done by the use of tables, graphs, and numbers (statistics) such as the mean, median, range, and standard deviation. This topic, **descriptive statistics**, is discussed in Chapter **??**.

- **Probability theory**, studied in Chapter **??**, is the theoretical tool of statistics. We saw in Example 1.2 how probability theory is used in the testing of hypotheses. There were no deaths in the treatment group, yet the evidence is not convincing that the Salk vaccine prevented any deaths. On the other hand, 22% of the 148 cases of paralytic polio fell into the treatment group, which is overwhelming evidence of the effectiveness of the vaccine in preventing paralytic polio.

- In Chapters **??**, **??**, **??**, **??**, and **??**, **probability models** are developed. We show that many of these models can be described in terms of standard population box models.[8] By translating story problems into probability box models, we can eliminate many of the extraneous details that complicate the analysis of case studies. We then develop procedures for analyzing the population models, whose solutions can be translated back into the story problems.

- In Chapter **??**, **sampling theory** is developed. It is essential to know the behavior of random samples from known populations before we do the inverse of making inferences about unknown populations by examining random samples.

- In Chapters **??**, **??**, and **??**, we will consider the statistical rules of evidence and how data can be used in **statistical testing of hypotheses**. We examine samples from unknown populations and, with the help of sampling theory, determine the plausibility of various hypotheses about the populations.

- Chapters **??** and **??** deal with the examination of data from unknown populations in order to make **estimates of parameters** of populations. We consider precision of estimates and the level of confidence of estimates.

- In Chapter **??**, we consider the **statistical relationship** of variables. In algebra, an equation such as $y = x^2$ describes a functional (or deterministic) relationship between x and y. Statistical relationships are not

[8] Population box models are, in turn, related to Polya Urn models as described in Marek Fisz, *Probability Theory and Mathematical Statistics*, third edition, John Wiley & Sons, Inc., 1963.

functional. For example, given a person's height x at age 14 and the same person's height y at age 21, the variables x and y are statistically (or stochastically) related, but not functionally related. **Correlation** measures the strength of the relationship of two statistical variables. Generally, tall 14-year-olds become tall 21-year-olds; and short 14-year-olds become short 21-year-olds, but you could not write a deterministic equation between x and y. We say that the two variables are positively correlated. Similarly, there is a negative correlation between the weight of a car and its fuel economy. We can use the strength of a relationship to predict the value of one variable from another. For example, if you had to predict the height of a 21-year-old person without knowing any other facts, your best guess is the average height for all 21-year-olds, plus or minus a standard error. However, if you know the person's height at age 14, along with other facts such as the correlation coefficient, the prediction can be greatly improved. The prediction will not be perfect, because the relationship is statistical. But the prediction will have more precision than one made without the knowledge of the height at age 14. The prediction of one variable from others is known as **regression analysis**.

- In Chapter **??**, we continue with statistical inference by comparing multiple dichotomous populations. We can test whether the historical data of a state lottery fit a model of equal probability for each of the possible outcomes. Or we can test whether any of a number of treatments makes a difference in the outcome of an illness.
- Finally, Chapter **??** covers resampling methods where we discuss parametric and nonparametric bootstrapping. This chapter is computer intensive and deals with inferences from small data sets.

Advice to the reader

At this point, it is a good idea to **review calculus**. It is not used heavily in the first week or so of the course but, beginning with Chapter **??**, you will need to know the basic facts of differentiation and integration.

REVIEW EXERCISES FOR CHAPTER 1

- **1.11** Cyberchondriacs are people who go online to search for information about health, medical care, or particular diseases. In a nationwide Harris Poll of 1001 adults surveyed between May 26 and June 10, 2000, 56% said that they

were online from home, office, school, library, or another location. Also, 86% of these 56% said that they have looked for health information online. This means that 48% of all American adults have looked for health information online. Based on the U.S. census bureau estimate of 204 million American adults, this amounts to 98 million people. In this situation, the number 56% is which of the following:

(a) a population (b) a sample (c) a parameter (d) a statistic

The number 98 million is an estimate of which of the above?

1.12 *The New York Times*/CBS News Poll uses a computer to randomly select phone numbers within each of the 42,000 residential exchanges in the country. This way the pollsters have access to both listed and unlisted numbers. In one survey, they called 1279 numbers. Of these, 352 were unlisted. This is not surprising because 29% of all numbers are unlisted. Select one option:

(a) 352 and 29 are both parameters;
(b) 352 and 29 are both statistics;
(c) 352 is a parameter and 29 is a statistic;
(d) 352 is a statistic and 29 is a parameter;
(e) none of the above.

1.13 (Proof) Recall from calculus the techniques of finding maximum values of functions.

a. Show that the maximum of the function $f(x) = x(1-x)$ occurs when $x = \frac{1}{2}$.

b. Use this to prove the inequality in Formula 1.1

$$\sqrt{\frac{\pi(1-\pi)}{n}} \leq \frac{1}{2\sqrt{n}}$$

1.14 Tossing a fair coin n times and counting heads is like n draws from a population box model consisting of cards numbered with 0's and 1's, and with a population percentage of $\pi = 0.50 = 50\%$. Suppose I toss a fair coin 100 times and compute the sample percentage p of heads. Use Formula 1.1 on page 6 to compute the standard error for the statistic p. Now, considering 400 tosses, what happens to SE_p?

• **1.15** Darius submits a computer-programming assignment. He claims that the program, when run on the busy university network, will have a run time to completion of 20 seconds, with a standard error of 5 seconds.

a. Assuming this to be true, and if his instructor were to test this program by running it twice, the total run time should be _____ seconds, give or take _____ seconds, on the average.

b. If the instructor runs Darius's program four times, the total run time should be _____, give or take _____, on the average.

1.16 Continue with the previous Exercise 1.15.

a. If the instructor were to execute the program a total of 100 times, the total run time should be _____, give or take _____, on the average.

b. Afterward, the instructor takes an average of the 100 runs. If the student's estimates are correct, the instructor should obtain an *average* run time of _____ seconds, give or take _____ seconds or so.

• **1.17** Consider walnuts packed in "one-pound" bags. Although the nominal weights of the bags are one pound, it's hard to put exactly one pound in each bag. As it turns out, the bags vary in weight by a standard error of approximately 1 ounce.

a. Use the Pythagorean property of standard errors to estimate the packaging error for the total weight of two packages.

b. What about the packaging error in a box of 16 bags?

⋆ **1.18** If the children who took part in the placebo control study of the Salk vaccine trials were assigned to the vaccine and placebo groups by chance, why weren't the two groups of the same size?

⋆ • **1.19** Why do you think so many children were needed for the Salk vaccine trials?

⋆ **1.20** Give an example of a study where a double-blind experiment is not possible.

* * * * * * * * * *

1.21 (Proof) Show that for two measurements, with random errors a and b, we have MA $\leq$ RMS; that is, show that the following is true for all $a \geq 0,\ b \geq 0$:

$$\frac{a+b}{2} \leq \sqrt{\frac{a^2+b^2}{2}}$$