INTRODUCTION TO BIOSTATISTICS

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 doesn't know, or can't 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 introduction we will give an overview of statistics along with an outline of the various topics in this course.

2 SAMPLING AND ESTIMATION

Harris Poll. Louis Harris and Associates (www.harrisinteractive.com) conduct polls on various topics, either face-to-face, by telephone, or by the internet. In one survey on health trends of adult Americans conducted in 1991 they contacted 1,256 randomly selected adults by phone and asked them questions about diet, stress management, seat belt use, etc. One of the questions asked was "Do you try hard to avoid too much fat in your diet?" They reported that 57% of the people responded YES to this question, which was a 2% increase from a similar survey conducted in 1983. 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 1,256 people. Such a subset of the population is called a **sample**.

Once every ten years the U.S. Census Bureau 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 of people's lives. We shall see as we progress through this course that, if done carefully, 1,256 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 one of several categories to each individual (such as person's blood type or hair color) it is called a **categorical variable**. And 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**.

A number derived from a *sample* is called a **statistic**, whereas a number derived from the *population* is called a **parameter**. Parameters are is 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.



Figure 1 Parameter and statistic for a dichotomous variable.

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 they say that the margin of error is plus or minus 3 percent, when such a small sample of all adult Americans were contacted? This is one of the questions that we will deal with in the course.

SIMULATION

Consider a box containing chips or cards, each of which is numbered either 0 or 1. We want to take a 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 don't 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:

Based on this 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 = .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, a computer program such as Minitab, or a table of random numbers. We will use Table 1 of random digits at the end of our textbook (on pages 638-9) to demonstrate the procedure. For convenience the random digits in the table are grouped into five digits. Each of the 50 rows holds 20 groups of fives. The entire table holds $50 \cdot 20 \cdot 5 = 5000$ random digits. To guard against using the same random numbers in every simulation we select to start in a random row and column, say row 16 column 06 (we can obtain a starting point by tossing a coin or paper clip on the page). The first 25 digits in row 16, starting with column 06, are

30144 29166 20915 53462 42573.

Since 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, resulting in the following values

11111 10100 11011 11101 11101.

In this simulation we ended up with 19 1's, which results in the statistic p = 19/25 = .76 = .76%. The random sampling error is +16%.

Continuing where we left off in row 16, the next 25 digits result in

01111 11101 10010 10101 01001.

This time we have 15 1's to obtain a statistic of p = 60% and a random sampling error of 0%.

One more time,

11111 01101 10111 00111 10110,

gives us 18 1's with p = 72%, and a random error of +12%.

The random errors were -4% from the box of cards and +16%, 0%, and +12% from the table of random numbers.

ERROR ANALYSIS

An **experiment** is a procedure which results in a measurement or observation. The Harris poll is an experiment which 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

AV:
$$\frac{-4\% + 16\% + 0\% + 12\%}{4} = +6\%.$$

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

MA:
$$\frac{|-4\%| + |+16\%| + |0\%| + |+12\%|}{4} = +8\%.$$

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

RMS:
$$\sqrt{\frac{(-4\%)^2 + (+16\%)^2 + (+0\%)^2 + (+12\%)^2}{4}} = \sqrt{120}\% = 10.2\%$$

The RMS is a more conservative measure of the typical size of the random sampling errors in the sense that $MA \leq RMS$.

For a given experiment the RMS of all possible random sampling errors is called the **standard error** (SE). We will more formally study the standard error later in the semester. For example, whenever we use a random sample of size n and its percentages p to estimate the population percentage π , we have

$$SE_p = \sqrt{\frac{\pi(1-\pi)}{n}} \le \frac{1}{2\sqrt{n}},\tag{1}$$

which for n = 25 comes out $SE = .097979 \cdots \leq .10$. Notice that the estimate $SE_p \leq 10\%$ is reasonably close to the RMS = 10.2% that we obtained in our four simulations. As the number of simulations increase, the RMS will tend to the SE. The *inequality* in (1) is useful because, in a situation such as the Harris poll, the value of the parameter π is generally unknown.

Notice from equation (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 larger, 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 about $\pm \frac{1}{2} \cdot 10\% = \pm 5\%$. If we combine our four samples in our simulations of size 25, we notice that we have a total of 14 + 13 + 19 + 12 = 58 cards numbered 1, which gives us p = 58% and thus a random error of -2%. 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 units, at this point we want to observe an interesting property. The standard errors of independent trials of a random experiments have a Pythagorean relationship, just as 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 $\sqrt{SE_A^2 + SE_B^2}$. In error analysis it is convenient to write

$$A = \alpha \pm SE_A,$$
$$B = \beta \pm SE_B,$$

and

$$A + B = \alpha + \beta \pm \sqrt{SE_A^2 + SE_B^2}.$$

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 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 about .4 miles. If A represents the distance measured on any given trip, we can write

 $A = 24.3 \pm .4$ miles.

This does not mean that *every* trip varies by .4 miles, rather it means that for such trips odometer readings average about 24.3 miles and vary by about .4 miles, or so.

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

 $B = 67.2 \pm 1.1$ 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

EXERCISES FOR SECTION 2

- 1. A medium apple has an average of about 80 calories. However, the number of calories A varies with the size and type of apple. Let's say that the standard error is about $SE_A = 10$ calories. Similarly the number of calories B in a bran muffin is about 400 calories with a standard error of about $SE_B = 40$ calories. Let C be the total number of calories in a lunch consisting of a medium apple and a bran muffin. Find the estimated total number of calories of C along with the standard error SE_C . Do the same for a lunch which also includes a cup of yogurt, which has 220 calories, give or take 15 calories, or so.
- 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 about 30 grams, give or take 5 grams, or so. The total weight T of the four mice will be about _____ grams, give or take _____ grams, or so. Use the Pythagorean property of standard errors.

3 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, about

95% of the time, the random error will be within this "margin in error." In the news media, the 95% condition is generally assumed but not always explicitly mentioned. We will show later in the course, that, for large random samples as in the Harris survey, this 95% margin of error is about 1.96 times (roughly *twice*) the standard error. Using the inequality in formula (1) we note that for n = 1,256, we obtain $SE_p \leq \frac{1}{2\sqrt{1256}} = .0141 = 1.41\%$, which is close to half of their reported 3%. Since 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 the estimate. For a given sample size, if we want to *increase* confidence, we have to *decrease* the precision, and *visa 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 later show that a margin of error of $\pm three$ standard errors, which gives us a precision of $\pm 3 \cdot \frac{1}{2\sqrt{1256}} = \pm .0423 = \pm 4.23\%$, will give us a statement with 99.7% confidence. And to get a statement with 99% confidence we need a margin of error of about ± 2.6 standard errors, which is about $\pm 3.6\%$. Using a margin of error of 1 standard error will result in a statement with only 68% confidence. At this point, don't worry about the specific percentages and levels of confidence. Details and derivations will come in later in the semester.

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 a 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 know 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 an experiment. Nonresponse bias is a problem encountered whenever a survey or census is attempted. Nonresponse rates are not uniform across all segments of a population, so that nonresponse could affect some groups more than others, biasing the results.

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

Measurement bias can also lead to errors. If a measurement procedure is not being applied correctly or consistently, all readings may be off.

Selection bias is another cause of error. If you haphazardly select lab mice for an experiment, generally you catch the slower one, the fatter ones, and the ones more friendly. If the remaining mice are used as controls, the two groups start the experiment of different vigor, weight, and temperament.

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.

EXERCISES FOR SECTION 3

- **3.** As part of a 1990 study on the causes of asthma, the parents of 939 seven-year old children in five German cities were interviewed¹. Of these children, the parents reported that 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.
- 4. The New York Times reported on a poll conducted October 18-21, 2000. This random phone survey found that among 1,010 registered votes, 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, choosing from the options

population, parameter, sample, statistic, unknown, π , p, 45%, 3%, 1.5%. The symbol for the parameter is _____. The value of 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 estimated to be _____. The 95% margin of error is estimated to be _____.

4 HYPOTHESIS TESTING

Salk Vaccine Trials. According to an article by Paul Meier², 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 the table 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 who had been given the salt solution.

¹Susanne Lau et al, Early exposure to house-dust mite and cat allergens and development of childhood asthma: a cohort study. The Lancet 356(2000), 1392-97.

²Paul Meier, The Biggest Public Health Experiment Ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine. This article appears in Statistics a Guide to the Unknown, Third Edition, Judith Tanur et al, editors, Duxbury Press, 1989.

	Number of	Fatal	Paralytic	Non-Paralytic	Total
Group	${f Subjects}$	Polio	& Fatal Polio	Polio	Cases
Vaccine	200,745	0	33	24	57
Placebo	201,229	4	115	27	142
TOTALS	401,974	4	148	51	199

Summary of Randomized Double Blind Salk Vaccine Trials.

Does the 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 about 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 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 is 50%, so the chance of four tails in sequence is 50% of 50% of 50% of 50%, or $(.50)^4 = .0625 = 6.25\%$. 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.

Now let us consider whether the vaccine was effective against paralytic polio. Of the children who were injected with the Salk vaccine 33 of them were later diagnosed with paralytic polio compared with 115 of the children in the placebo group. Using a similar analysis, we suppose 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 coin come up 33 heads and 115 tails. In 148 tosses you expect about half of them to be heads; that is, about 74 should be heads. 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.

5 PREVIEW

Here are some of the issues that we will consider this semester.

- How do you take a sample? What is a random sample? Unless you carefully design an experiment, the data collected may be subject to bias. In the Harris study, how do you account for people that can't be reached by phone, don't 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**. Chapter 8 of the textbook deals with this topic. However, the issues involved are pointed out throughout the book.
- 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 is called **descriptive statistics**, which is discussed in Chapter 2 of the textbook.
- **Probability theory**, studied in Chapter 3 and 4 of the textbook, is the theoretical tool of statistics. We saw in Example 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.
- Sampling theory is in Chapter 5 of the text. It is essential to know the behavior of random samples from known populations before we do the inverse of making inferences about unknown populations through the examination of random samples.
- Chapter 6 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 7 and 9 we will consider the statistical rules of evidence and how data can be used in **statistical test of hypotheses**. We examine samples from unknown populations and, with the help of sampling theory, we determine the plausibility of various hypotheses about the populations.
- Chapters 10 and 11 is a continuation of estimation and hypothesis testing, except that we deal with multiple comparisons of treatments.
- In Chapter 12 we study 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 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 couldn't write an 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, you could use a population box model. This would result in a prediction consisting of the mean of the population box, or 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**.

• Chapter 13 is merely an overview of all of the statistical methods presented in the text.

6 REVIEW EXERCISES

- 5. Cyberchondriacs are people who go on-line to search for information about health, medical care, or particular diseases. In a nationwide Harris Poll of 1,001 adults, surveyed between May 26 and June 10, 2000, 56% said that they were on-line from home, office, school, library, or other location. Also 86% of these 56% said that they have looked for health information on-line. This means that 48% of all American adults have looked for health information on-line. Based on a the U.S. Census Bureau estimate of 204 million American adults, this amounts to about 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.
- 6. 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 about 1 ounce. Use the Pythagorean property of standard errors to estimate the packaging error for the total weight of two packages. What about the packaging error in a box of 16 bags?
- 7. 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 the same size?
- 8. Why do you think so many children were needed for the Salk vaccine trials?
- 9. Tossing a coin n times and counting heads is like n draws from a population box model consisting of cards numbered with zeros and ones, and with a population percentage of $\pi = .50 = 50\%$. Suppose I toss a coin 100 times and compute the sample percentage p of heads. Use formula (1) to compute the standard error for the statistic p. Now considering 400 tosses, what happens to SE_p ?