Introduction to Statistics

During the next hour, while you are reading this chapter, 250 Americans will die, and the chances are one in a million that you'll be one of them. Don't stop reading, however, since that won't reduce the probability. In fact, putting your book down might even increase the probability, especially if you decide to leave your room and go outside to engage in some high-risk activity.

According to the TV special "Against All Odds," if you go rock climbing, the probability of your getting killed is 200 in a million; or parachuting, 250 in a million; or hang gliding, 1140 in a million. So, sit safely still and read on, and while doing that let's look at some more of life's probabilities: the probability of having your car stolen this year, 1 out of 120; of a pregnant woman having twins, 1 out of 90 (or triplets, 1 out of 8000); of a young adult (18–22) being paroled from prison and then being rearrested for a serious crime, 7 out of 10; and of any single American baby becoming a genius (IQ of 135 or higher), less than 1 out of 100 (Krantz, 1992). Incidentally, by the time you finish reading Chapter 6, you'll be able to calculate that genius probability value—even if you're not a genius yourself.

As you probably know, most accidents occur at home, since typical Americans spend most of their time there. And 25% of all home accidents occur in the bathroom—falling in the tub, getting cut while shaving, and so forth. Don't assume, however, that you'll be safe if you decide to shave in the kitchen instead. Also, we can predict with a high degree of accuracy that during the next year 9000 pedestrians will be killed by a moving car. But this statistic does not tell us which 9000. Understanding probability situations is an important aspect of life, so maybe there are some good reasons for getting involved in statistical thinking.

Rather than continuing to list reasons why you should take a first course in statistics, let's assume that it is probably a required course and that you have to take it anyway. Perhaps you have put it off for quite a while, until there is no choice left but to "bite the bullet" and get it over with. This is not to say that all of you have been dragged, kicking and screaming, into this course; however, as statisticians would put it, the probability is high that this hypothesis is true for some of you.

Graphs and Measures of Central Tendency

To be able to describe large amounts and, of course, small amounts of data, statisticians have created a series of tools or abbreviated symbols to give structure and meaning to the apparent chaos of the original measures. Thousands, even millions, of scores can be organized and neatly summarized by the appropriate use of descriptive techniques. These summaries allow for precise communication of whatever story the data have to tell.

During this initial foray into the realm of descriptive statistics, we shall concentrate on two major techniques: graphs and measures of central tendency.

Variability

Just as measures of central tendency give information about the similarity among scores, ***measures of variability*** *give information about how scores differ or vary. Usually when a group of persons, things, or events are measured, the measurements are scattered over a range of values. Sometimes the scatter is large, as in the case of a distribution of the incomes of all wage-earning Americans. Some people earn almost nothing; others earn millions of dollars. Sometimes the scatter is small, as it is among the heights of women in a Las Vegas chorus line. In this case, the shortest and tallest dancer may differ by only an inch or two. The fact that measures of people vary describes the concept of "individual differences," the theme running through all of the social sciences. The description of data is never complete until some indication of the variability is found.*

For example, suppose we are told that the yearly mean temperature in Anchorage, Alaska, is 56°F, and that in Honolulu, Hawaii, the mean is only 8 degrees higher, 64°F. If this mean temperature comparison were the only information you were given about the two areas, you might be led to believe that the climates are fairly similar. Once some of the variability facts are added, however, you begin to realize how misleading it can be to compare averages alone. In Honolulu, the temperature rarely rises about 84°F or dips below 50°F. In Anchorage, on the other hand, it may reach 100°F during the summer (midnight sun time) and often drops to –40°F during the winter. It is possible that the temperature in Anchorage hits the mean of 56°F only twice a year, once in the spring as the temperatures are going up and once in the fall as they are coming down. Obviously, to interpret the measures of central tendency correctly, we must know something about variability.

The Normal Curve and z Scores

If the mean and the standard deviation are the heart and soul of descriptive statistics, then the normal curve is its lifeblood. Although there is some dispute among statisticians as to when and by whom the normal curve was first introduced, it is customary to credit its discovery to the great German mathematician **Karl Friedrich Gauss.** *Even to this day, many statisticians refer to the normal curve as the Gaussian curve.*

As a child, Gauss was one of those "perfect-pitch" mathematical prodigies who put most students and teachers in a state of total terror. It is said that he could add, subtract, multiply, and divide before he could talk. It has never been clear how he communicated this facility to his family and friends. Did he, like the Wonder Horse, stamp out his answers with his foot? We are also told that at the age of 3, when he presumably could talk, he detected a math error on his father's pay envelope. When he was 8, he startled his schoolmaster by adding all the numbers from 1 through 100 in his head. (The teacher had given the class this busy-work assignment to do longhand while he corrected papers. Obviously, little Karl's lightning calculation changed the teacher's plans.)

The Hypothesis of Association: Correlation

We now enter the treacherous and murky waters of correlation. Perhaps no other area of research demands more caution or is fraught with more danger. Too many of us assume that because the correlation is easy to say, that it's likewise easy to understand. In some respects, it is. Mathematically, the correlation coefficient is rather straightforward and easily calculated. What can legitimately be inferred from its numerical value, however, is quite another story.

The meaning of the word correlation comes literally from its parts: "co" means with, together, or jointly, and "relation" means association. Thus, when two events regularly occur together, then they are said to be correlated, as with blond hair and blue eyes. Also, when changes in one set of events are regularly accompanied by changes in another set of events, correlation is said to exist; for example, as children get taller, they also tend to get heavier. As we shall soon see, the correlation coefficient provides us with a numerical value that states the extent to which two events, or two sets of measurements, tend to occur or to change together—the extent to which they covary.

Parameter Estimates and Hypothesis Testing

As we have seen, knowing the mean and the standard error of the sampling distribution of means is of critical importance. These values allow us to use the z score table, which in turn permits us to make probability statements regarding where specific samples might fall. We have also seen, however, that to obtain the mean and the standard deviation of this important distribution, we must select every last sample and measure each individual in all the samples—in short, we must measure the entire population. Whenever we do so, the resulting values are the parameters. When the parameters are known, there is nothing left to predict—there is no need for inferential statistics.

The job of the statistician, then, is to estimate these important parameters, to predict their values without measuring the entire population. This is done by measuring a random sample, calculating the resulting values, called statistics, and using only these statistics for inferring the parameters.

The Hypothesis of Difference

Up to this point, we have dealt exclusively with samples selected one at a time. Just as we previously learned to make probability statements about where individual raw scores might fall, we next learned to make similar probability statements regarding where the mean of a specific sample might fall. Note the emphasis of part of the definition of the estimated standard error of the mean: "an estimate made on the basis of information contained in a single sample." In many research situations, however, we must select more than one sample. Any time the researcher wants to compare *sample groups to find out, for example, if one sample is quicker, or taller, or wealthier than another sample, it is obvious that at least two sample groups must be selected.*

Compared to What?

When the social philosopher and one-line comic Henny Youngman was asked how his wife was, he always replied, "Compared to what?" (Youngman, 1991). The researcher must constantly ask and then answer the same question. In Chapter 1, we pointed to the fallacious example used to argue that capital punishment does not deter crime. The example cited was that when pickpockets were publicly hanged, other pickpockets were on hand to steal from the watching crowd. Since pockets were picked at the hanging, so goes the argument, obviously hanging does not deter pickpocketing. Perhaps this is true, but we can still ask, "Compared to what?" Compared to the number of pockets picked at less grisly public gatherings? For such an observation of behavior to have any meaning, then, a control or comparison group is needed. We cannot say that Sample A is different from Sample B if, in fact, there is no Sample B. In this chapter, then, the focus will be on making comparisons between pairs *of sample means. The underlying aim of the entire chapter is to discover the logical concepts involved in selecting a pair of samples and then to determine whether or not these samples can be said to represent a single population.*