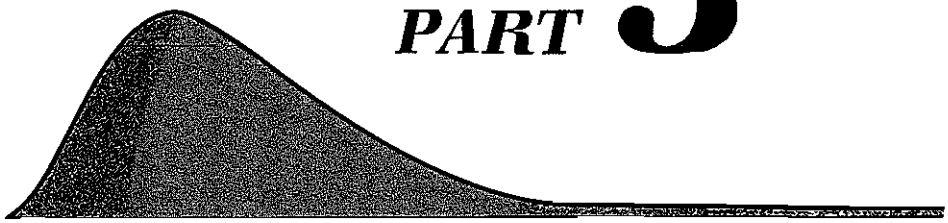
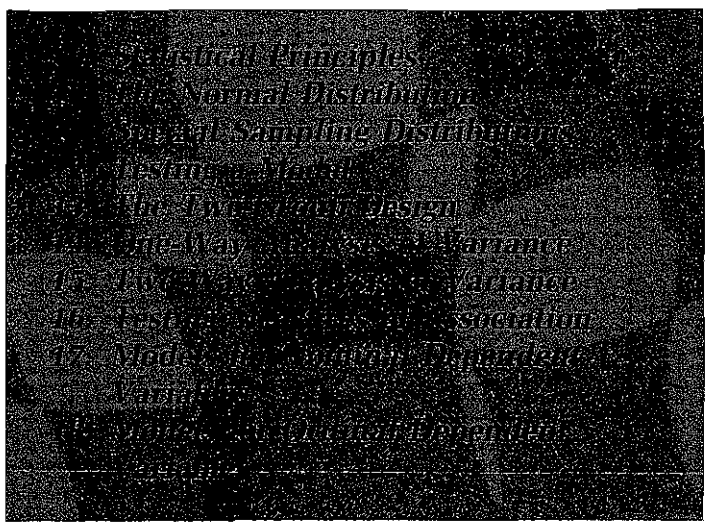
PART 3



Inferential Statistics



9 Statistical Principles

In the previous eight chapters many topics have been covered. Formulas for means, variances, regression, and correlation coefficients have been presented. Methods have been presented to *describe* in rather rich detail a sample of numbers. *Descriptive statistics* are used to produce quantitative summaries of numbers.

Yet there is more to data analysis than just describing the data at hand. Besides describing the data, statistics are used to test ideas, theories, and hypotheses about the population. The testing of hypotheses, or more generally the testing of models, is called *inferential statistics*.

In this chapter and the next two, the groundwork is prepared for inferential statistics. There are many important statistical concepts that are essential to the understanding how models are tested. The chapter begins with a discussion of the way in which numbers are chosen to form a sample. Then, the idea that a statistic has a distribution is presented. The next topic concerns criteria that are used to determine which statistic is best. The final topic is the binomial distribution.

All of the topics and concepts in this chapter relate to statistical theory. Many of the quantities referred to in this chapter are computed from hypothetical distributions. Though somewhat abstract, the ideas in this chapter are essential for the intelligent comprehension of the remainder of the book.

Sample and Population

When we step on the scale to check our weight, we care about the number that appears. Our own numbers or scores are important to us. However, we are generally not interested in the numbers or scores of particular other persons. Similarly, we do not care if John's IQ is 123 or whether Paul shocked the person in the Milgram experiment at 450 volts. For a sample of numbers, the specific numbers by themselves are ordinarily of little interest. They become interesting if they are viewed as representative of the numbers that some

larger group of persons would provide. The magic of statistics is that the numbers of a few can be used to gauge the numbers of many. If a researcher takes an election survey and finds that a candidate is losing by 5%, the candidate does not really care that he or she lost in the survey. It is the views of all the voters that is important, but a survey of a few persons may reflect the views of many voters.

This process of going from the few to the many is valid only if a set of traditional assumptions are true: One assumption is that there is a set of objects called the *population* and a number is attached to each object. It is also assumed that the set of objects is infinite in size. From the population a small set of observations called the *sample* is gathered. Thus the sample is actually a subset of all the possible observations. The population is infinite in size, whereas the sample is finite.

Although classical statistical theory is based on the assumption that the population is infinite in size, in practice most populations in the social and behavioral sciences are finite in size. For instance, in an election survey the population is the set of potential voters, which is finite in size. It has been found that the theory based on infinite populations can be safely applied to large but finite populations.

The optimal size for a sample depends on several factors. Sometimes a small sample of ten objects is plenty, whereas in other cases hundreds of objects are needed. The choice of sample size depends on the statistic being computed, the resources available, and the degree of confidence required in the conclusions.

Sample data are used to infer properties of the population. For instance, the mean of a sample is computed and used to infer the mean of the population. A quantity computed from the sample is called a *statistic*. So the mean, variance, and correlation coefficient when computed from sample data are called statistics. A quantity computed using all the members of a population is called a *parameter*. The population mean and variance are parameters. Roman letters $\bar{X}$, s, and r are used to designate statistics, while Greek letters μ (mu), σ (sigma), and ρ (rho) are used to designate parameters.

The value of a parameter is almost never known and, as a result, it remains a hypothetical value that is estimated by a statistic. It would be nearly impossible, for example, to interview all the voters before an election. So from the population a subset or a sample of voters is selected. A major part of statistical theory concerns how sample data can be used to describe accurately the population from which they are drawn. Ideally the sample should correspond as closely to the population as possible so that the statistics computed from sample data will be as close to the population parameters as possible. However, it is impossible for a finite sample to mirror an infinite population exactly. So whenever one samples from the population, there will be *sampling errors*. Sampling and sampling error go hand in hand.

The term *sampling error* is unfortunate because an error seems to imply a

mistake that could have been prevented. Sampling error cannot be prevented. It is necessarily built into sampling. When a few observations are used to represent the many, errors necessarily follow.

Imagine working for a company that produces a certain type of machine that is ordered by more than 1000 factories throughout the world. You want to project how many of your machines will be ordered during the upcoming year. It would be wasteful of time and money to contact managers at the 1000 factories and ask each how many of the product they plan to order for the next year. You cannot afford to interview every member of the population of 1000 factories. Practical considerations force you to choose only a *sample* of factories. Any statistic computed from the sample would not be the same as the population parameter. Thus your estimate of demand will not be what it would be if you surveyed all the 1000 factories. In essence, your survey may be in error. At issue are the procedures that can reduce the amount of sampling error and procedures for determining the likely amount of sampling error. It must be realized that sampling error is inherent in these procedures. So the goal must be to minimize and measure it.

Random and Independent Sampling

In order to reduce the amount of sampling error and to estimate its probable extent, one must sample from the population in certain prespecified ways. Sampling must be random and independent. Random and independent sampling does not prevent sampling errors. It merely reduces their size and permits the determination of the likely amount of sampling error.

There are two strategies for controlling the amount of sampling error. First, the objects are selected randomly. *Random sampling* requires that every object from the population is equally likely to be chosen to become a member of the sample. Second, the objects are chosen independently. *Independent sampling* requires that if a given object is chosen, it in no way increases or decreases the probability that any other object is subsequently chosen. So for a voter survey, persons are randomly chosen. One should not sample haphazardly by interviewing persons who happen to pass by on a busy street and are willing to be interviewed. A "grab sample" of friends and acquaintances is not a random sample.

Random and independent sampling is the cornerstone of classical sampling theory. There are two major ways to obtain a random sample. The first way is to use a method that makes the decision of which objects or persons to include in the sample random. For instance, one can flip a coin or roll a die to make a decision randomly. The second procedure to achieve random sampling from a population of human subjects is to use a random number table. Such a table is presented in Appendix B.

To use a random number table, first a list of the population members is

made. Next, these persons are assigned sequential numbers from 1 to N, where N is the population size. Assume that the number N has k digits. A number is picked from a random number table, not necessarily the first number in the table. From this random number the first k digits are examined, and it is determined whether anyone in the population has the number. If someone does, that person is included in the sample. The first k digits of the next random number are used to select the next person into the sample. This process is repeated until the desired sample size is achieved.

Independence requires that the probability of a person being sampled not change if anyone else is sampled. For instance, if the population is a city, using all the persons living on one block violates the independent sampling requirement even if the block is chosen at random: If one person is sampled, the person's next-door neighbor must be sampled. However, if one die is repeatedly rolled, a sample of rolls is independent. Rolling a six on one trial does not change the probability of rolling a six on the next trial. The inferential statistical methods discussed in the subsequent chapters are based on the assumption that the data were gathered by means of independent and random sampling methods. The major way that the independent sampling assumption is violated is by measuring a person more than once.

When sampling from a population, one can sample with or without replacement. When sampling without replacement, once an object is chosen to be a member of the sample, it cannot be chosen again. When sampling with replacement, an object can be chosen again. With infinite populations, there is no practical difference between sampling with and without replacement. For small populations, observations can be independent only if one samples with replacement. If two cards are sampled without replacement from a deck of cards, the sampling is not independent. Picking an ace as the first card decreases the probability of picking an ace as the second card if that first ace is not returned to the deck.

There are two major reasons why objects should be randomly and independently sampled. The first is to reduce the amount of sampling error. In the absence of any other information, random and independent sampling provides statistics that are as close to the parameter as possible. The second reason is that random and independent sampling permits the quantification of the amount of sampling error. Thus, it is known how close, in theory, the statistic is to the population parameter.

In reality samples used in social and the behavioral sciences are not randomly or independently formed. For instance, in most psychology experiments students sign up to serve as subjects. Such samples are not random. Moreover, subjects when they are sampled randomly are hardly ever sampled with replacement. (Only in survey research, such as election surveys, are persons randomly and independently sampled.) Although persons are not randomly sampled, it can be argued that the response from a person is a

random sample of that subject's behavior. So the responses can be assumed to be a random sample of a population of responses.

Sampling Distribution

A statistic is a number that is computed from a sample. If the same statistic were computed from a different sample taken from the same population, almost certainly a different result would be obtained and neither would exactly equal the population parameter. This variation from sample to sample is called sampling error. For instance, if the average height in a class of 20 students were measured from a random sample of 5 persons, the mean of this sample of 5 will almost certainly not equal the mean of another sample of 5. In any given sample, the people chosen will be a bit taller or shorter than the class average. Sampling error goes hand in hand with statistical estimation. A sampling error is not an intentional mistake; rather, it is an inevitable outcome. In other words, error in the estimation of population parameters is the inevitable price that must be paid for the ease and economy afforded by sampling.

The sampling distribution of a statistic can be conceptualized as follows: If the mean were computed from two different samples with the same n drawn from the same population, two different values would be obtained. If an infinite number of samples of size n were drawn and for each the mean were computed, a frequency distribution of the sample means of size n could be created. The distribution of this infinite set of means is referred to as the *random sampling distribution* of the mean, or more simply as the *sampling distribution* of the mean. In general, any statistic has a sampling distribution. A *sampling distribution of a statistic* is a theoretical distribution based on an imaginary repeated sampling and computation of a statistic.

A sampling distribution has two important properties: its mean and its standard deviation. The mean of the sampling distribution equals what the statistic tends to be on the average. The standard deviation describes how variable the statistic is when repeatedly calculated from different samples of the same size. So, the mean of the sampling distribution states what the statistic is estimating and the standard deviation states how close it comes to that value on the average. The standard deviation of the random sampling distribution of a statistic is called the *standard error of the statistic*. The standard error quantifies the amount of sampling error in the statistic. It measures the degree to which the statistic would be likely to change if another sample were drawn and the statistic were recomputed. In other words, the standard error of a statistic measures how variable a statistic is when it is recomputed using a different sample of the same size.

For most statistics, the standard error decreases as the sample size in-

creases. In fact, if a statistic did not have this attribute, it would be deemed a poor statistic. The relation between sample size and the standard error can be illustrated for the sample mean. If the population variance is 100, the standard error of the sample mean takes on the following values for the given sample sizes:

n	*Standard Error of* $\bar{X}$
10	3.16
25	2.00
100	1.00
150	.82

(The exact formula for the standard error of the mean is presented in Chapter 11.) The standard error of 3.16 for the sample size of 10 implies that the typical difference of the sample mean from the population mean is 3.16 units. When the n is 150, the sample mean differs from the population mean by about .82 unit. Normally the effect of increasing the sample size suffers from the "law of diminishing returns." Doubling the sample size does not cut the standard error in half. To cut the standard error of the mean in half, the sample size must be quadrupled. This can be seen in the above table: The standard error for a sample of 25 is twice that for a sample of 100 subjects.

As the sample size increases, the statistic does not vary as much from sample to sample. However, even with large sample sizes, the sample statistic still does not exactly equal the population parameter. Sampling and sampling error go hand in hand. The standard error quantifies the amount of sampling error. The standard error states how close the statistic is to the parameter, on the average, not in any particular instance. So if an election survey shows that a candidate is leading an election by 8% and the standard error is only 5%, it does not guarantee that the candidate is ahead. The 5% standard error is the average or typical amount of error to be expected given the sample size. The actual error in a particular survey may be zero or even 20%. The error in any particular study is never known. Only known is the average or standard error across many studies of the same sample size.

Properties of Statistics

In computing measures of central tendencies various issues arose. For instance, as was explained in Chapter 3, there are three measures of central tendency: the mean, the median, and the mode. Is there any reason to prefer one measure over the other? Also the variance is divided by $n - 1$ and not n. Why is this? These questions can be answered once two important properties of statistics are defined.

Bias

The mean of the sampling distribution can be used to define an important property of a statistic—namely, bias. A statistic that is supposed to estimate a population parameter is said to be an *unbiased* estimate of that parameter if the mean of the random sampling distribution of the statistic equals the parameter. Unbiased statistics exactly estimate on the average what they purport to be estimating. If a statistic is unbiased, the statistic itself does not necessarily equal the parameter; it only does so on the average. A positively biased estimate statistic is one that tends, on average, to overestimate the parameter value.

The sample mean $\bar{X}$ is an unbiased estimate of the population mean, μ. If the distribution is symmetric, the median is also an unbiased estimate of the population mean. If the distribution is symmetric, the mean is also an unbiased estimate of the population median.

The sample variance s^2 is an unbiased estimate of the population variance, σ^2. The formula for s^2 has $n - 1$ in the denominator and not n. The reason for this is to make s^2 an unbiased estimator of σ^2. If the denominator of s^2 were n instead of $n - 1$, s^2 would be a biased estimator.

A fact not very well known is that the sample correlation coefficient r is a slightly biased estimate of the population correlation when the population correlation ρ is nonzero. For sample sizes of five or more the sample correlation coefficient slightly underestimates positive values of ρ and overestimates negative values of ρ. For moderate and large samples, the bias is trivially small and can be safely ignored. Although the correlation coefficient is biased, the regression coefficient is not.

Efficiency

A second important property of a statistic is efficiency. One statistic is said to be relatively more *efficient* than another statistic if its standard error is smaller than that of the other statistic. Thus, if one statistic's standard error is smaller than another's, the former is said to be more efficient than the latter.

To choose between two statistics, say the mean and the median, their efficiency must be considered. Which of the two statistics is more efficient depends on the shape of the distribution. Thus it is necessary to consider what the underlying distribution is before determining which statistic is most efficient.

As will be discussed in the next chapter, in data analysis it is generally assumed that the population distribution is normal. For this reason, the discussion of efficiency will presume that the population distribution is normal.

When the population distribution is normal, $\bar{X}$ and s^2 are unbiased es-

timators of μ and σ^2, respectively. Moreover, they are the most efficient unbiased estimators of the parameters, again when the distribution is normal.

Although the sample mean and variance are optimal when the distributions are normal, they can be very inefficient when outliers are present. In particular, the sample standard deviation is a very inefficient estimate of the population standard deviation when there are outliers in the sample. A statistic whose efficiency is not affected much by nonnormality or outliers is said to be *robust*. The median is more robust than the mean, and the interquartile range is more robust than the standard deviation.

Binomial Distribution

One important population distribution is the binomial distribution, which can be used to describe a series of random events. Before getting into the mathematics of this distribution, consider the following example.

Imagine someone flipping a coin four times and counting the number of heads. That number can vary from zero, no heads at all, to all four flips being heads. One might wonder what the probability is of obtaining exactly three heads in four flips. To determine this and other probabilities, the binomial distribution is used.

First, the probability of flipping a head on a single flip must be determined. That probability is .5 and is denoted as p. The probability of not flipping heads must be $1 - p$, which is denoted as q. So the probability of a success (flipping heads) is denoted as p, and the probability of a failure (flipping tails) is denoted as q.

Second, it must be assumed that the trials (flips) are independent. That is, having flipped a heads in trial one does not increase or decrease the probability of flipping a heads in trial two. It seems reasonable to believe that coin flips are in fact independent. However, other events are not. If a trial is picking a card from a deck, then the chances of picking an ace are 1/13 or .077. If the card from trial one is not replaced, then the chances of picking an ace on trial two are affected by what happened on trial one. If an ace were picked on trial one, then the chances of picking an ace in trial two are 1/17 or .059. But if some other card besides an ace were picked in trial one, then the chances of picking an ace in trial two are 4/51 or .078.

If there are a set of independent trials with a known probability, the probability of a given outcome can be determined. Assuming that there are n trials, the probability of x successes, given a binomial distribution, is:

$$\frac{n!}{x!(n-x)!}\,p^x q^{n-x}$$

The term $n!$ is read as n factorial. It equals

$$n(n-1)(n-2)\ldots(3)(2)(1)$$

So 5! equals (5)(4)(3)(2)(1) = 120. By convention 0! = 1.

Using the binomial formula, the probability of obtaining three heads in four flips can now be computed. The terms for the binomial formula are

p(the probability of a success) = .5
q(the probability of a failure) = 1 − .5 = .5
n(the number of trials) = 4
x(the number of successes) = 3

Putting the terms in the formula yields:

$$\frac{4!}{3!\ 1!}.5^3.5^1 = \frac{(4)(3)(2)(1)}{(3)(2)(1)(1)}(.125)(.5) = .25$$

So the probability of flipping three heads in four trials is .25.

As a second example, consider the probability of a subject getting eight of ten answers correct on a recognition test with five alternatives if the subject is just guessing. The probability of a correct guess on each trial is 1/5 or .20. So the terms for the formula are

p(the probability of a success) = .20
q(the probability of a failure) = 1 − .20 = .80
n(the number of trials) = 10
x(the number of successes) = 8

Putting these terms into the binomial formula yields:

$$\frac{10!}{8!\ 2!}.20^8.80^2 = .0000737$$

If a variable has a binomial distribution, its population mean equals np and its population variance equals npq. So for the prior example with $n = 10$, $p = .20$, and $q = .80$, the mean is 2.0 and the variance is 1.6. These are *population parameters*. The mean and variance of sample of subjects would not exactly equal these hypothetical values because of sampling error.

Summary

Descriptive statistics are used to summarize the scores in a sample. Examples of descriptive statistics are the mean and the standard deviation. *Inferential statistics* are used to test models.

A *population* is an infinite set of objects and the *sample* is a subset of objects chosen from the population. A *statistic* is a number, like the mean or variance, computed from sample data. A *parameter* is a number computed from all the possible values of the population. The variation of a statistic from sample to sample is called *sampling error*.

A *random* sample is a set of numbers chosen so that each object is equally likely to be chosen to be a member of the sample. Sampling is said to be

independent if choosing one object in no way changes the probability that some other object will be chosen. Random sampling can be accomplished by rolling a die or using a random number table. Random and independent sampling minimizes the amount of sampling error. Sampling can either be with or without replacement. When sampling *without replacement,* once the object is sampled, it cannot be sampled again. When sampling *with replacement,* an object can be sampled again.

The same statistic could be computed from many different samples drawn from the same population. A statistic has a *random sampling distribution* or, more simply, a *sampling distribution.* The standard deviation of the sampling distribution is called the *standard error* of the statistic.

A statistic is *unbiased* if the mean of its random sampling distribution is equal to the parameter that it is supposed to be estimating. Both $\bar{X}$ and s^2 are unbiased statistics. One statistic is more *efficient* than another if the variance of its random sampling distribution is less than the other statistic's variance. A statistic is said to be *robust* if an outlier in the sample does not dramatically change the value of the statistic. The median is a more robust estimate of the population mean than the sample mean. The interquartile range is a more robust estimator of variability than the sample variance.

The binomial distribution is used to describe the probability of an event happening x times in n trials. The probability of a success must be known, and trials must be independent.

Problems

1. Which of the following schemes are random samples of pages from the phone book?

   a. page 10 through 53
   b. every fifth page
   c. opening the book 50 times and picking a page
   d. using a random number table to pick 60 pages

2. Imagine two statistics p and q that are both estimators of the same parameter. Presume that p and q are estimated from a sample of size 50. Also presume that p is unbiased and its standard error for an n of 50 is .88. The estimate q is biased tending to be .01 unit too high and its standard error is .44. Which of the two statistics would you prefer to use and why?

3. Define in words each of the following for the variable of reaction time.

   a. a sample mean of 555 milliseconds (ms)
   b. a population mean of 531 ms
   c. a sample standard deviation of 112 ms

d. a population standard deviation of 123 ms
e. a standard error of the mean (based on ten observations) of 12.3 ms

4. Indicate whether each of the following statements is true or false.

   a. The mean of a random sample equals the mean of the population.
   b. The mean of a random sample will tend to equal the mean of the population.

5. Imagine two statistics k and q that both estimate parameter theta. Assume that the sampling distribution of k for $n = 100$ has a mean of theta and a variance of 20, and the sampling distribution of q for $n = 100$ has a mean of theta and a variance of 25.

   a. Are k and q unbiased estimates of theta?
   b. Which is more efficient, k or q?
   c. What are the standard errors of k and q?

6. Given that the population correlation ρ equals zero, the quantity $r/\sqrt{(1 - r^2)}$ has a sampling distribution with a mean of zero and a variance of approximately $1/(n - 2)$.

   a. Is $r/\sqrt{(1 - r^2)}$ an unbiased estimate of $\rho/\sqrt{(1 - \rho^2)}$?
   b. What is the standard error of $r/\sqrt{(1 - r^2)}$?

7. For a flat distribution why is the sample median an unbiased estimate of the population mean?

8. If the population is normally distributed, it can be shown that the squared interquartile range multiplied by .55 is essentially an unbiased estimate of the population variance. What estimate of σ^2 would you prefer: s^2 or the adjusted interquartile range statistic? Why?

9. An unbiased estimate of the population mean is any randomly sampled observation in the sample. Why is the sample mean preferable to the single-score estimate of the mean?

10. Explain why each of the following is not an independent sample of college students.

    a. all students in one randomly chosen dormitory
    b. students waiting in line for a movie

11. For the population of 9, 8, 12, and 6, the following are all 16 samples with replacement of sample size 2.

    (9, 9), (9, 8) (9, 12), (9, 6)

    (8, 9), (8, 8) (8, 12), (8, 6)

(12, 9), (12, 8) (12, 12), (12, 6)

(6, 9), (6, 8), (6, 12), (6, 6)

Compute from each $\bar{X}$ and make a frequency table of the random sampling distribution of $\bar{X}$. Compute the mean and standard deviation of the sampling distribution.

12. Imagine that a random sample of 50 out of 12,938 students from a university is needed. State how a random sample could be drawn using a random number table.

13. Imagine a sample of three numbers from a population. An estimate, called U, of the population mean is

$$.5X_1 + .25X_2 + .25X_3$$

It turns out that U is an unbiased estimate of the population mean with a standard error of $.612\sigma$ where σ is the population standard deviation. The sample mean has a standard error of $.577\sigma$. What statistic is more efficient: $\bar{X}$ or U? Why?

14. Which of the following can be considered a random and independent sample of students from a classroom?

a. All students whose names begin with A and K
b. All students who sit in the first row.
c. All students who volunteer to be in a study.
d. The first ten students who come to class one day.

15. What is the probability of rolling a single die six times and obtaining a five three times?

16. If one rolls two dice, the probability of rolling an eight is 5/36. What then is the probability of rolling an eight on four out of five rolls?

17. Jim feels down on his luck. He has bet a number ten times straight on roulette and lost. The chances of hitting the number are 1 out of 38. What is the probability of losing ten times in a row?

18. If the probability of getting divorced is .42, what is the probability that in a sample of seven, five couples will get divorced?