Chapter 8

Fundamental Sampling Distributions and Data Descriptions

8.1 Random Sampling

The outcome of a statistical experiment may be recorded either as a numerical value or as a descriptive representation. When a pair of dice is tossed and the total is the outcome of interest, we record a numerical value. However, if the students of a certain school are given blood tests and the type of blood is of interest, then a descriptive representation might be more useful. A person's blood can be classified in 8 ways: AB, A, B, or O, each with a plus or minus sign, depending on the presence or absence of the Rh antigen.

In this chapter, we focus on sampling from distributions or populations and study such important quantities as the *sample mean* and *sample variance*, which will be of vital importance in future chapters. In addition, we attempt to give the reader an introduction to the role that the sample mean and variance will play in statistical inference in later chapters. The use of modern high-speed computers allows the scientist or engineer to greatly enhance his or her use of formal statistical inference with graphical techniques. Much of the time, formal inference appears quite dry and perhaps even abstract to the practitioner or to the manager who wishes to let statistical analysis be a guide to decision-making.

Populations and Samples

We begin this section by discussing the notions of *populations* and *samples*. Both are mentioned in a broad fashion in Chapter 1. However, much more needs to be presented about them here, particularly in the context of the concept of random variables. The totality of observations with which we are concerned, whether their number be finite or infinite, constitutes what we call a **population**. There was a time when the word *population* referred to observations obtained from statistical studies about people. Today, statisticians use the term to refer to observations relevant to anything of interest, whether it be groups of people, animals, or all possible outcomes from some complicated biological or engineering system.

Definition 8.1: A **population** consists of the totality of the observations with which we are concerned.

The number of observations in the population is defined to be the size of the population. If there are 600 students in the school whom we classified according to blood type, we say that we have a population of size 600. The numbers on the cards in a deck, the heights of residents in a certain city, and the lengths of fish in a particular lake are examples of populations with finite size. In each case, the total number of observations is a finite number. The observations obtained by measuring the atmospheric pressure every day, from the past on into the future, or all measurements of the depth of a lake, from any conceivable position, are examples of populations whose sizes are infinite. Some finite populations are so large that in theory we assume them to be infinite. This is true in the case of the population of lifetimes of a certain type of storage battery being manufactured for mass distribution throughout the country.

Each observation in a population is a value of a random variable X having some probability distribution $f(x)$. If one is inspecting items coming off an assembly line for defects, then each observation in the population might be a value 0 or 1 of the Bernoulli random variable X with probability distribution

$$b(x; 1, p) = p^x q^{1-x}, \qquad x = 0, 1$$

where 0 indicates a nondefective item and 1 indicates a defective item. Of course, it is assumed that p, the probability of any item being defective, remains constant from trial to trial. In the blood-type experiment, the random variable X represents the type of blood and is assumed to take on values from 1 to 8. Each student is given one of the values of the discrete random variable. The lives of the storage batteries are values assumed by a continuous random variable having perhaps a normal distribution. When we refer hereafter to a "binomial population," a "normal population," or, in general, the "population $f(x)$," we shall mean a population whose observations are values of a random variable having a binomial distribution, a normal distribution, or the probability distribution $f(x)$. Hence, the mean and variance of a random variable or probability distribution are also referred to as the mean and variance of the corresponding population.

In the field of statistical inference, statisticians are interested in arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population. For example, in attempting to determine the average length of life of a certain brand of light bulb, it would be impossible to test all such bulbs if we are to have any left to sell. Exorbitant costs can also be a prohibitive factor in studying an entire population. Therefore, we must depend on a subset of observations from the population to help us make inferences concerning that same population. This brings us to consider the notion of sampling.

Definition 8.2: A **sample** is a subset of a population.

If our inferences from the sample to the population are to be valid, we must obtain samples that are representative of the population. All too often we are

tempted to choose a sample by selecting the most convenient members of the population. Such a procedure may lead to erroneous inferences concerning the population. Any sampling procedure that produces inferences that consistently overestimate or consistently underestimate some characteristic of the population is said to be **biased**. To eliminate any possibility of bias in the sampling procedure, it is desirable to choose a **random sample** in the sense that the observations are made independently and at random.

In selecting a random sample of size n from a population $f(x)$, let us define the random variable X_i, $i = 1, 2, \ldots, n$, to represent the ith measurement or sample value that we observe. The random variables $X_1, X_2, \ldots, X_n$ will then constitute a random sample from the population $f(x)$ with numerical values $x_1, x_2, \ldots, x_n$ if the measurements are obtained by repeating the experiment n independent times under essentially the same conditions. Because of the identical conditions under which the elements of the sample are selected, it is reasonable to assume that the n random variables $X_1, X_2, \ldots, X_n$ are independent and that each has the same probability distribution $f(x)$. That is, the probability distributions of $X_1, X_2, \ldots, X_n$ are, respectively, $f(x_1), f(x_2), \ldots, f(x_n)$, and their joint probability distribution is $f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$. The concept of a random sample is described formally by the following definition.

Definition 8.3: Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having the same probability distribution $f(x)$. Define $X_1, X_2, \ldots, X_n$ to be a **random sample** of size n from the population $f(x)$ and write its joint probability distribution as

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

If one makes a random selection of $n = 8$ storage batteries from a manufacturing process that has maintained the same specification throughout and records the length of life for each battery, with the first measurement x_1 being a value of X_1, the second measurement x_2 a value of X_2, and so forth, then $x_1, x_2, \ldots, x_8$ are the values of the random sample $X_1, X_2, \ldots, X_8$. If we assume the population of battery lives to be normal, the possible values of any X_i, $i = 1, 2, \ldots, 8$, will be precisely the same as those in the original population, and hence X_i has the same identical normal distribution as X.

8.2 Some Important Statistics

Our main purpose in selecting random samples is to elicit information about the unknown population parameters. Suppose, for example, that we wish to arrive at a conclusion concerning the proportion of coffee-drinkers in the United States who prefer a certain brand of coffee. It would be impossible to question every coffee-drinking American in order to compute the value of the parameter p representing the population proportion. Instead, a large random sample is selected and the proportion $\hat{p}$ of people in this sample favoring the brand of coffee in question is calculated. The value $\hat{p}$ is now used to make an inference concerning the true proportion p.

Now, $\hat{p}$ is a function of the observed values in the random sample; since many

random samples are possible from the same population, we would expect $\hat{p}$ to vary somewhat from sample to sample. That is, $\hat{p}$ is a value of a random variable that we represent by P. Such a random variable is called a **statistic**.

Definition 8.4: Any function of the random variables constituting a random sample is called a **statistic**.

Location Measures of a Sample: The Sample Mean, Median, and Mode

In Chapter 4 we introduced the two parameters μ and σ^2, which measure the center of location and the variability of a probability distribution. These are constant population parameters and are in no way affected or influenced by the observations of a random sample. We shall, however, define some important statistics that describe corresponding measures of a random sample. The most commonly used statistics for measuring the center of a set of data, arranged in order of magnitude, are the **mean**, **median**, and **mode**. Although the first two of these statistics were defined in Chapter 1, we repeat the definitions here. Let $X_1, X_2, \ldots, X_n$ represent n random variables.

(a) Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Note that the statistic $\bar{X}$ assumes the value $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ when X_1 assumes the value x_1, X_2 assumes the value x_2, and so forth. The term *sample mean* is applied to both the statistic $\bar{X}$ and its computed value $\bar{x}$.

(b) Sample median:

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$$

The sample median is also a location measure that shows the middle value of the sample. Examples for both the sample mean and the sample median can be found in Section 1.3. The sample mode is defined as follows.

(c) The sample mode is the value of the sample that occurs most often.

Example 8.1: Suppose a data set consists of the following observations:

0.32 0.53 0.28 0.37 0.47 0.43 0.36 0.42 0.38 0.43.

The sample mode is 0.43, since this value occurs more than any other value. ◢

As we suggested in Chapter 1, a measure of location or central tendency in a sample does not by itself give a clear indication of the nature of the sample. Thus, a measure of variability in the sample must also be considered.

Variability Measures of a Sample: The Sample Variance, Standard Deviation, and Range

The variability in a sample displays how the observations spread out from the average. The reader is referred to Chapter 1 for more discussion. It is possible to have two sets of observations with the same mean or median that differ considerably in the variability of their measurements about the average.

Consider the following measurements, in liters, for two samples of orange juice bottled by companies A and B:

Sample A	0.97	1.00	0.94	1.03	1.06
Sample B	1.06	1.01	0.88	0.91	1.14

Both samples have the same mean, 1.00 liter. It is obvious that company A bottles orange juice with a more uniform content than company B. We say that the **variability**, or the **dispersion**, of the observations from the average is less for sample A than for sample B. Therefore, in buying orange juice, we would feel more confident that the bottle we select will be close to the advertised average if we buy from company A.

In Chapter 1 we introduced several measures of sample variability, including the **sample variance**, **sample standard deviation**, and **sample range**. In this chapter, we will focus mainly on the sample variance. Again, let $X_1, \ldots, X_n$ represent n random variables.

(a) Sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{8.2.1}$$

The computed value of S^2 for a given sample is denoted by s^2. Note that S^2 is essentially defined to be the average of the squares of the deviations of the observations from their mean. The reason for using $n-1$ as a divisor rather than the more obvious choice n will become apparent in Chapter 9.

Example 8.2: A comparison of coffee prices at 4 randomly selected grocery stores in San Diego showed increases from the previous month of 12, 15, 17, and 20 cents for a 1-pound bag. Find the variance of this random sample of price increases.

***Solution*:** Calculating the sample mean, we get

$$\bar{x} = \frac{12 + 15 + 17 + 20}{4} = 16 \text{ cents}.$$

Therefore,

$$s^2 = \frac{1}{3}\sum_{i=1}^{4}(x_i - 16)^2 = \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3}$$
$$= \frac{(-4)^2 + (-1)^2 + (1)^2 + (4)^2}{3} = \frac{34}{3}.$$

Whereas the expression for the sample variance best illustrates that S^2 is a measure of variability, an alternative expression does have some merit and thus the reader should be aware of it. The following theorem contains this expression.

Theorem 8.1: If S^2 is the variance of a random sample of size n, we may write

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right].$$

***Proof*:** By definition,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2)$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right].$$

As in Chapter 1, the **sample standard deviation** and the **sample range** are defined below.

(b) Sample standard deviation:

$$S = \sqrt{S^2},$$

where S^2 is the sample variance.

Let $X_{\max}$ denote the largest of the X_i values and $X_{\min}$ the smallest.

(c) Sample range:

$$R = X_{\max} - X_{\min}.$$

Example 8.3: Find the variance of the data 3, 4, 5, 6, 6, and 7, representing the number of trout caught by a random sample of 6 fishermen on June 19, 1996, at Lake Muskoka.

***Solution*:** We find that $\sum_{i=1}^{6} x_i^2 = 171$, $\sum_{i=1}^{6} x_i = 31$, and $n = 6$. Hence,

$$s^2 = \frac{1}{(6)(5)}[(6)(171) - (31)^2] = \frac{13}{6}.$$

Thus, the sample standard deviation $s = \sqrt{13/6} = 1.47$ and the sample range is $7 - 3 = 4$.

Exercises

8.1 Define suitable populations from which the following samples are selected:

(a) Persons in 200 homes in the city of Richmond are called on the phone and asked to name the candidate they favor for election to the school board.

(b) A coin is tossed 100 times and 34 tails are recorded.

(c) Two hundred pairs of a new type of tennis shoe were tested on the professional tour and, on average, lasted 4 months.

(d) On five different occasions it took a lawyer 21, 26, 24, 22, and 21 minutes to drive from her suburban home to her midtown office.

8.2 The lengths of time, in minutes, that 10 patients waited in a doctor's office before receiving treatment were recorded as follows: 5, 11, 9, 5, 10, 15, 6, 10, 5, and 10. Treating the data as a random sample, find

(a) the mean;

(b) the median;

(c) the mode.

8.3 The reaction times for a random sample of 9 subjects to a stimulant were recorded as 2.5, 3.6, 3.1, 4.3, 2.9. 2.3, 2.6, 4.1, and 3.4 seconds. Calculate

(a) the mean;

(b) the median.

8.4 The number of tickets issued for traffic violations by 8 state troopers during the Memorial Day weekend are 5, 4, 7, 7, 6, 3, 8, and 6.

(a) If these values represent the number of tickets issued by a random sample of 8 state troopers from Montgomery County in Virginia, define a suitable population.

(b) If the values represent the number of tickets issued by a random sample of 8 state troopers from South Carolina, define a suitable population.

8.5 The numbers of incorrect answers on a true-false competency test for a random sample of 15 students were recorded as follows: 2, 1, 3, 0, 1, 3, 6, 0, 3, 3, 5, 2, 1, 4, and 2. Find

(a) the mean;

(b) the median;

(c) the mode.

8.6 Find the mean, median, and mode for the sample whose observations, 15, 7, 8, 95, 19, 12, 8, 22, and 14, represent the number of sick days claimed on 9 federal income tax returns. Which value appears to be the best measure of the center of these data? State reasons for your preference.

8.7 A random sample of employees from a local manufacturing plant pledged the following donations, in dollars, to the United Fund: 100, 40, 75, 15, 20, 100, 75, 50, 30, 10, 55, 75, 25, 50, 90, 80, 15, 25, 45, and 100. Calculate

(a) the mean;

(b) the mode.

8.8 According to ecology writer Jacqueline Killeen, phosphates contained in household detergents pass right through our sewer systems, causing lakes to turn into swamps that eventually dry up into deserts. The following data show the amount of phosphates per load of laundry, in grams, for a random sample of various types of detergents used according to the prescribed directions:

Laundry Detergent	**Phosphates per Load (grams)**
A & P Blue Sail	48
Dash	47
Concentrated All	42
Cold Water All	42
Breeze	41
Oxydol	34
Ajax	31
Sears	30
Fab	29
Cold Power	29
Bold	29
Rinso	26

For the given phosphate data, find

(a) the mean;

(b) the median;

(c) the mode.

8.9 Consider the data in Exercise 8.2, find

(a) the range;

(b) the standard deviation.

8.10 For the sample of reaction times in Exercise 8.3, calculate

(a) the range;

(b) the variance, using the formula of form (8.2.1).

8.11 For the data of Exercise 8.5, calculate the variance using the formula

(a) of form (8.2.1);

(b) in Theorem 8.1.

8.12 The tar contents of 8 brands of cigarettes selected at random from the latest list released by the Federal Trade Commission are as follows: 7.3, 8.6, 10.4, 16.1, 12.2, 15.1, 14.5, and 9.3 milligrams. Calculate

(a) the mean;

(b) the variance.

8.13 The grade-point averages of 20 college seniors selected at random from a graduating class are as follows:

3.2	1.9	2.7	2.4	2.8
2.9	3.8	3.0	2.5	3.3
1.8	2.5	3.7	2.8	2.0
3.2	2.3	2.1	2.5	1.9

Calculate the standard deviation.

8.14 (a) Show that the sample variance is unchanged if a constant c is added to or subtracted from each

value in the sample.

(b) Show that the sample variance becomes c^2 times its original value if each observation in the sample is multiplied by c.

8.15 Verify that the variance of the sample 4, 9, 3, 6, 4, and 7 is 5.1, and using this fact, along with the results of Exercise 8.14, find

(a) the variance of the sample 12, 27, 9, 18, 12, and 21;

(b) the variance of the sample 9, 14, 8, 11, 9, and 12.

8.16 In the 2004-05 football season, University of Southern California had the following score differences for the 13 games it played.

11 49 32 3 6 38 38 30 8 40 31 5 36

Find

(a) the mean score difference;

(b) the median score difference.

8.3 Sampling Distributions

The field of statistical inference is basically concerned with generalizations and predictions. For example, we might claim, based on the opinions of several people interviewed on the street, that in a forthcoming election 60% of the eligible voters in the city of Detroit favor a certain candidate. In this case, we are dealing with a random sample of opinions from a very large finite population. As a second illustration we might state that the average cost to build a residence in Charleston, South Carolina, is between $330,000 and $335,000, based on the estimates of 3 contractors selected at random from the 30 now building in this city. The population being sampled here is again finite but very small. Finally, let us consider a soft-drink machine designed to dispense, on average, 240 milliliters per drink. A company official who computes the mean of 40 drinks obtains $\bar{x} = 236$ milliliters and, on the basis of this value, decides that the machine is still dispensing drinks with an average content of $\mu = 240$ milliliters. The 40 drinks represent a sample from the infinite population of possible drinks that will be dispensed by this machine.

Inference about the Population from Sample Information

In each of the examples above, we computed a statistic from a sample selected from the population, and from this statistic we made various statements concerning the values of population parameters that may or may not be true. The company official made the decision that the soft-drink machine dispenses drinks with an average content of 240 milliliters, even though the sample mean was 236 milliliters, because he knows from sampling theory that, if $\mu = 240$ milliliters, such a sample value could easily occur. In fact, if he ran similar tests, say every hour, he would expect the values of the statistic $\bar{x}$ to fluctuate above and below $\mu = 240$ milliliters. Only when the value of $\bar{x}$ is substantially different from 240 milliliters will the company official initiate action to adjust the machine.

Since a statistic is a random variable that depends only on the observed sample, it must have a probability distribution.

Definition 8.5: The probability distribution of a statistic is called a **sampling distribution**.

The sampling distribution of a statistic depends on the distribution of the population, the size of the samples, and the method of choosing the samples. In the

remainder of this chapter we study several of the important sampling distributions of frequently used statistics. Applications of these sampling distributions to problems of statistical inference are considered throughout most of the remaining chapters. The probability distribution of $\bar{X}$ is called the **sampling distribution of the mean**.

What Is the Sampling Distribution of $\bar{X}$?

We should view the sampling distributions of $\bar{X}$ and S^2 as the mechanisms from which we will be able to make inferences on the parameters μ and σ^2. The sampling distribution of $\bar{X}$ with sample size n is the distribution that results when an **experiment is conducted over and over** (always with sample size n) **and the many values of $\bar{X}$ result**. This sampling distribution, then, describes the variability of sample averages around the population mean μ. In the case of the soft-drink machine, knowledge of the sampling distribution of $\bar{X}$ arms the analyst with the knowledge of a "typical" discrepancy between an observed $\bar{x}$ value and true μ. The same principle applies in the case of the distribution of S^2. The sampling distribution produces information about the variability of s^2 values around σ^2 in repeated experiments.

8.4 Sampling Distribution of Means and the Central Limit Theorem

The first important sampling distribution to be considered is that of the mean $\bar{X}$. Suppose that a random sample of n observations is taken from a normal population with mean μ and variance σ^2. Each observation X_i, $i = 1, 2, \ldots, n$, of the random sample will then have the same normal distribution as the population being sampled. Hence, by the reproductive property of the normal distribution established in Theorem 7.11, we conclude that

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

has a normal distribution with mean

$$\mu_{\bar{X}} = \frac{1}{n}(\underbrace{\mu + \mu + \cdots + \mu}_{n \text{ terms}}) = \mu \text{ and variance } \sigma^2_{\bar{X}} = \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ terms}}) = \frac{\sigma^2}{n}.$$

If we are sampling from a population with unknown distribution, either finite or infinite, the sampling distribution of $\bar{X}$ will still be approximately normal with mean μ and variance σ^2/n, provided that the sample size is large. This amazing result is an immediate consequence of the following theorem, called the Central Limit Theorem.

The Central Limit Theorem

Theorem 8.2: **Central Limit Theorem:** If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

The normal approximation for $\bar{X}$ will generally be good if $n \geq 30$, provided the population distribution is not terribly skewed. If $n < 30$, the approximation is good only if the population is not too different from a normal distribution and, as stated above, if the population is known to be normal, the sampling distribution of $\bar{X}$ will follow a normal distribution exactly, no matter how small the size of the samples.

The sample size $n = 30$ is a guideline to use for the Central Limit Theorem. However, as the statement of the theorem implies, the presumption of normality on the distribution of $\bar{X}$ becomes more accurate as n grows larger. In fact, Figure 8.1 illustrates how the theorem works. It shows how the distribution of $\bar{X}$ becomes closer to normal as n grows larger, beginning with the clearly nonsymmetric distribution of an individual observation ($n = 1$). It also illustrates that the mean of $\bar{X}$ remains μ for any sample size and the variance of $\bar{X}$ gets smaller as n increases.

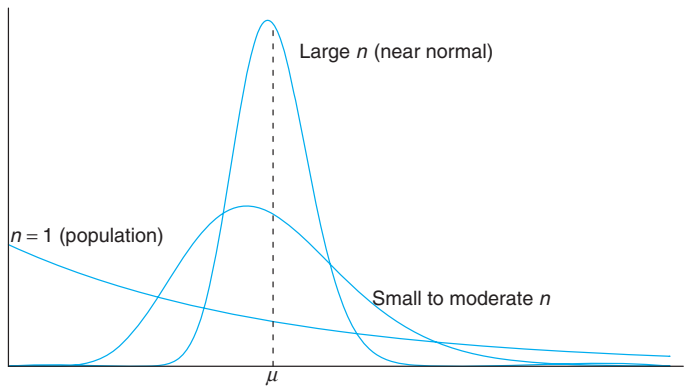

Figure 8.1: Illustration of the Central Limit Theorem (distribution of $\bar{X}$ for $n = 1$, moderate n, and large n).

Example 8.4: An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

***Solution*:** The sampling distribution of $\bar{X}$ will be approximately normal, with $\mu_{\bar{X}} = 800$ and $\sigma_{\bar{X}} = 40/\sqrt{16} = 10$. The desired probability is given by the area of the shaded

region in Figure 8.2.

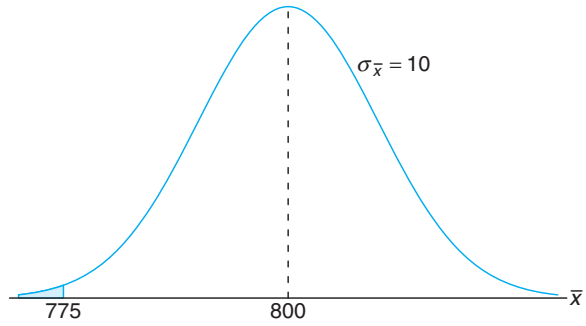

Figure 8.2: Area for Example 8.4.

Corresponding to $\bar{x} = 775$, we find that

$$z = \frac{775 - 800}{10} = -2.5,$$

and therefore

$$P(\bar{X} < 775) = P(Z < -2.5) = 0.0062.$$

Inferences on the Population Mean

One very important application of the Central Limit Theorem is the determination of reasonable values of the population mean μ. Topics such as hypothesis testing, estimation, quality control, and many others make use of the Central Limit Theorem. The following example illustrates the use of the Central Limit Theorem with regard to its relationship with μ, the mean of the population, although the formal application to the foregoing topics is relegated to future chapters.

In the following case study, an illustration is given which draws an inference that makes use of the sampling distribution of $\bar{X}$. In this simple illustration, μ and σ are both known. The Central Limit Theorem and the general notion of sampling distributions are often used to produce evidence about some important aspect of a distribution such as a parameter of the distribution. In the case of the Central Limit Theorem, the parameter of interest is the mean μ. The inference made concerning μ may take one of many forms. Often there is a desire on the part of the analyst that the data (in the form of $\bar{x}$) support (or not) some predetermined conjecture concerning the value of μ. The use of what we know about the sampling distribution can contribute to answering this type of question. In the following case study, the concept of hypothesis testing leads to a formal objective that we will highlight in future chapters.

Case Study 8.1: **Automobile Parts**:An important manufacturing process produces cylindrical component parts for the automotive industry. It is important that the process produce

parts having a mean diameter of 5.0 millimeters. The engineer involved conjectures that the population mean is 5.0 millimeters. An experiment is conducted in which 100 parts produced by the process are selected randomly and the diameter measured on each. It is known that the population standard deviation is $\sigma = 0.1$ millimeter. The experiment indicates a sample average diameter of $\bar{x} = 5.027$ millimeters. Does this sample information appear to support or refute the engineer's conjecture?

***Solution*:** This example reflects the kind of problem often posed and solved with hypothesis testing machinery introduced in future chapters. We will not use the formality associated with hypothesis testing here, but we will illustrate the principles and logic used.

Whether the data support or refute the conjecture depends on the probability that data similar to those obtained in this experiment ($\bar{x} = 5.027$) can readily occur when in fact $\mu = 5.0$ (Figure 8.3). In other words, how likely is it that one can obtain $\bar{x} \geq 5.027$ with $n = 100$ if the population mean is $\mu = 5.0$? If this probability suggests that $\bar{x} = 5.027$ is not unreasonable, the conjecture is not refuted. If the probability is quite low, one can certainly argue that the data do not support the conjecture that $\mu = 5.0$. The probability that we choose to compute is given by $P(|\bar{X} - 5| \geq 0.027)$.

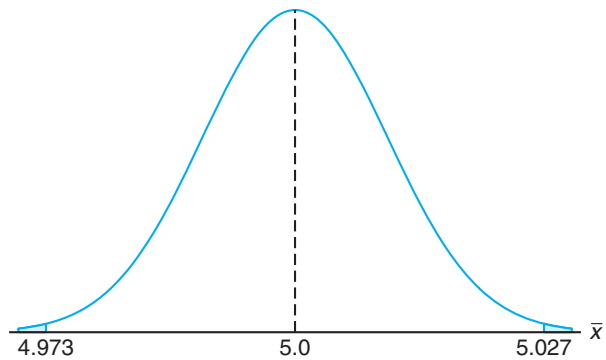

Figure 8.3: Area for Case Study 8.1.

In other words, if the mean μ is 5, what is the chance that $\bar{X}$ will deviate by as much as 0.027 millimeter?

$$P(|\bar{X} - 5| \geq 0.027) = P(\bar{X} - 5 \geq 0.027) + P(\bar{X} - 5 \leq -0.027)$$
$$= 2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq 2.7\right).$$

Here we are simply standardizing $\bar{X}$ according to the Central Limit Theorem. If the conjecture $\mu = 5.0$ is true, $\frac{\bar{X}-5}{0.1/\sqrt{100}}$ should follow $N(0, 1)$. Thus,

$$2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq 2.7\right) = 2P(Z \geq 2.7) = 2(0.0035) = 0.007.$$

Therefore, one would experience by chance that an $\bar{x}$ would be 0.027 millimeter from the mean in only 7 in 1000 experiments. As a result, this experiment with $\bar{x} = 5.027$ certainly does not give supporting evidence to the conjecture that $\mu = 5.0$. In fact, it strongly refutes the conjecture!

Example 8.5: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute.

***Solution*:** In this case, $\mu = 28$ and $\sigma = 3$. We need to calculate the probability $P(\bar{X} > 30)$ with $n = 40$. Since the time is measured on a continuous scale to the nearest minute, an $\bar{x}$ greater than 30 is equivalent to $\bar{x} \geq 30.5$. Hence,

$$P(\bar{X} > 30) = P\left(\frac{\bar{X} - 28}{5/\sqrt{40}} \geq \frac{30.5 - 28}{5/\sqrt{40}}\right) = P(Z \geq 3.16) = 0.0008.$$

There is only a slight chance that the average time of one bus trip will exceed 30 minutes. An illustrative graph is shown in Figure 8.4.

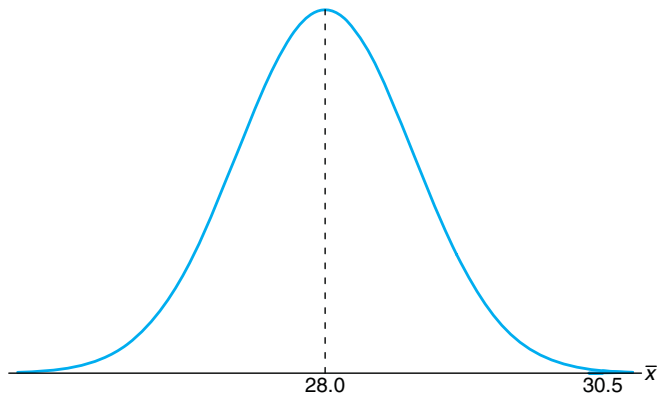

Figure 8.4: Area for Example 8.5.

Sampling Distribution of the Difference between Two Means

The illustration in Case Study 8.1 deals with notions of statistical inference on a single mean μ. The engineer was interested in supporting a conjecture regarding a single population mean. A far more important application involves two populations. A scientist or engineer may be interested in a comparative experiment in which two manufacturing methods, 1 and 2, are to be compared. The basis for that comparison is $\mu_1 - \mu_2$, the difference in the population means.

Suppose that we have two populations, the first with mean μ_1 and variance σ_1^2, and the second with mean μ_2 and variance σ_2^2. Let the statistic $\bar{X}_1$ represent the mean of a random sample of size n_1 selected from the first population, and the statistic $\bar{X}_2$ represent the mean of a random sample of size n_2 selected from

the second population, independent of the sample from the first population. What can we say about the sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$ for repeated samples of size n_1 and n_2? According to Theorem 8.2, the variables $\bar{X}_1$ and $\bar{X}_2$ are both approximately normally distributed with means μ_1 and μ_2 and variances σ_1^2/n_1 and σ_2^2/n_2, respectively. This approximation improves as n_1 and n_2 increase. By choosing independent samples from the two populations we ensure that the variables $\bar{X}_1$ and $\bar{X}_2$ will be independent, and then using Theorem 7.11, with $a_1 = 1$ and $a_2 = -1$, we can conclude that $\bar{X}_1 - \bar{X}_2$ is approximately normally distributed with mean

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

and variance

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

The Central Limit Theorem can be easily extended to the two-sample, two-population case.

Theorem 8.3: If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

is approximately a standard normal variable.

If both n_1 and n_2 are greater than or equal to 30, the normal approximation for the distribution of $\bar{X}_1 - \bar{X}_2$ is very good when the underlying distributions are not too far away from normal. However, even when n_1 and n_2 are less than 30, the normal approximation is reasonably good except when the populations are decidedly nonnormal. Of course, if both populations are normal, then $\bar{X}_1 - \bar{X}_2$ has a normal distribution no matter what the sizes of n_1 and n_2 are.

The utility of the sampling distribution of the difference between two sample averages is very similar to that described in Case Study 8.1 on page 235 for the case of a single mean. Case Study 8.2 that follows focuses on the use of the difference between two sample means to support (or not) the conjecture that two population means are the same.

Case Study 8.2: **Paint Drying Time**: Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.

Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A$ and $\bar{X}_B$ are average drying times for samples of size $n_A = n_B = 18$.

***Solution*:** From the sampling distribution of $\bar{X}_A - \bar{X}_B$, we know that the distribution is approximately normal with mean

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_A - \mu_B = 0$$

and variance

$$\sigma^2_{\bar{X}_A - \bar{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}.$$

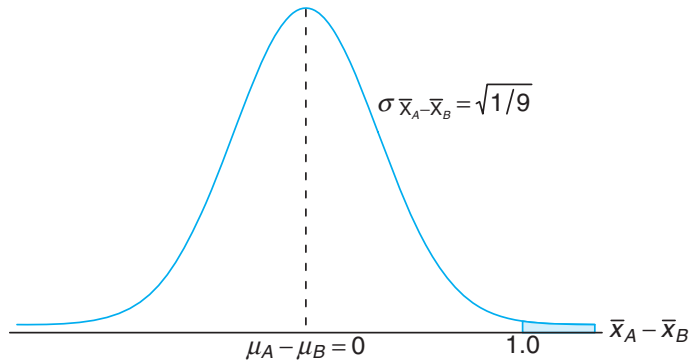

Figure 8.5: Area for Case Study 8.2.

The desired probability is given by the shaded region in Figure 8.5. Corresponding to the value $\bar{X}_A - \bar{X}_B = 1.0$, we have

$$z = \frac{1 - (\mu_A - \mu_B)}{\sqrt{1/9}} = \frac{1 - 0}{\sqrt{1/9}} = 3.0;$$

so

$$P(Z > 3.0) = 1 - P(Z < 3.0) = 1 - 0.9987 = 0.0013.$$

What Do We Learn from Case Study 8.2?

The machinery in the calculation is based on the presumption that $\mu_A = \mu_B$. Suppose, however, that the experiment is actually conducted for the purpose of drawing an inference regarding the equality of μ_A and μ_B, the two population mean drying times. If the two averages differ by as much as 1 hour (or more), this clearly is evidence that would lead one to conclude that the population mean drying time is not equal for the two types of paint. On the other hand, suppose

that the difference in the two sample averages is as small as, say, 15 minutes. If $\mu_A = \mu_B$,

$$P[(\bar{X}_A - \bar{X}_B) > 0.25 \text{ hour}] = P\left(\frac{\bar{X}_A - \bar{X}_B - 0}{\sqrt{1/9}} > \frac{3}{4}\right)$$
$$= P\left(Z > \frac{3}{4}\right) = 1 - P(Z < 0.75) = 1 - 0.7734 = 0.2266.$$

Since this probability is not low, one would conclude that a difference in sample means of 15 minutes can happen by chance (i.e., it happens frequently even though $\mu_A = \mu_B$). As a result, that type of difference in average drying times certainly *is not a clear signal* that $\mu_A \neq \mu_B$.

As we indicated earlier, a more detailed formalism regarding this and other types of statistical inference (e.g., hypothesis testing) will be supplied in future chapters. The Central Limit Theorem and sampling distributions discussed in the next three sections will also play a vital role.

Example 8.6: The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 year, while those of manufacturer B have a mean lifetime of 6.0 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer A will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B?

***Solution*:** We are given the following information:

Population 1	Population 2
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

If we use Theorem 8.3, the sampling distribution of $\bar{X}_1 - \bar{X}_2$ will be approximately normal and will have a mean and standard deviation

$$\mu_{\bar{X}_1 - \bar{X}_2} = 6.5 - 6.0 = 0.5 \quad \text{and} \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} = 0.189.$$

The probability that the mean lifetime for 36 tubes from manufacturer A will be at least 1 year longer than the mean lifetime for 49 tubes from manufacturer B is given by the area of the shaded region in Figure 8.6. Corresponding to the value $\bar{x}_1 - \bar{x}_2 = 1.0$, we find that

$$z = \frac{1.0 - 0.5}{0.189} = 2.65,$$

and hence

$$P(\bar{X}_1 - \bar{X}_2 \geq 1.0) = P(Z > 2.65) = 1 - P(Z < 2.65)$$
$$= 1 - 0.9960 = 0.0040.$$

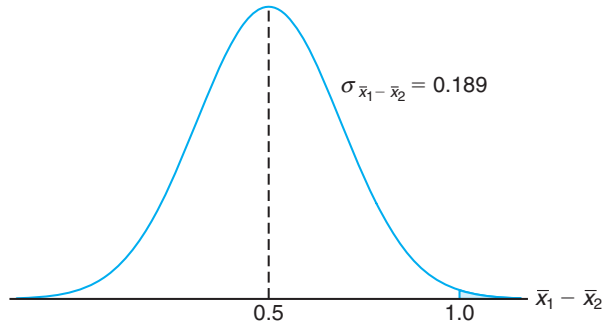

Figure 8.6: Area for Example 8.6.

More on Sampling Distribution of Means—Normal Approximation to the Binomial Distribution

Section 6.5 presented the normal approximation to the binomial distribution at length. Conditions were given on the parameters n and p for which the distribution of a binomial random variable can be approximated by the normal distribution. Examples and exercises reflected the importance of the concept of the "normal approximation." It turns out that the Central Limit Theorem sheds even more light on how and why this approximation works. We certainly know that a binomial random variable is the number X of successes in n independent trials, where the outcome of each trial is binary. We also illustrated in Chapter 1 that the proportion computed in such an experiment is an average of a set of 0s and 1s. Indeed, while the proportion X/n is an average, X is the sum of this set of 0s and 1s, and both X and X/n are approximately normal if n is sufficiently large. Of course, from what we learned in Chapter 6, we know that there are conditions on n and p that affect the quality of the approximation, namely $np \geq 5$ and $nq \geq 5$.

Exercises

8.17 If all possible samples of size 16 are drawn from a normal population with mean equal to 50 and standard deviation equal to 5, what is the probability that a sample mean $\bar{X}$ will fall in the interval from $\mu_{\bar{X}} - 1.9\sigma_{\bar{X}}$ to $\mu_{\bar{X}} - 0.4\sigma_{\bar{X}}$? Assume that the sample means can be measured to any degree of accuracy.

8.18 If the standard deviation of the mean for the sampling distribution of random samples of size 36 from a large or infinite population is 2, how large must the sample size become if the standard deviation is to be reduced to 1.2?

8.19 A certain type of thread is manufactured with a mean tensile strength of 78.3 kilograms and a standard deviation of 5.6 kilograms. How is the variance of the sample mean changed when the sample size is
(a) increased from 64 to 196?
(b) decreased from 784 to 49?

8.20 Given the discrete uniform population

$$f(x) = \begin{cases} \frac{1}{3}, & x = 2, 4, 6, \\ 0, & \text{elsewhere,} \end{cases}$$

find the probability that a random sample of size 54, selected with replacement, will yield a sample mean greater than 4.1 but less than 4.4. Assume the means are measured to the nearest tenth.

8.21 A soft-drink machine is regulated so that the amount of drink dispensed averages 240 milliliters with

a standard deviation of 15 milliliters. Periodically, the machine is checked by taking a sample of 40 drinks and computing the average content. If the mean of the 40 drinks is a value within the interval $\mu_{\bar{X}} \pm 2\sigma_{\bar{X}}$, the machine is thought to be operating satisfactorily; otherwise, adjustments are made. In Section 8.3, the company official found the mean of 40 drinks to be $\bar{x} = 236$ milliliters and concluded that the machine needed no adjustment. Was this a reasonable decision?

8.22 The heights of 1000 students are approximately normally distributed with a mean of 174.5 centimeters and a standard deviation of 6.9 centimeters. Suppose 200 random samples of size 25 are drawn from this population and the means recorded to the nearest tenth of a centimeter. Determine

(a) the mean and standard deviation of the sampling distribution of $\bar{X}$;

(b) the number of sample means that fall between 172.5 and 175.8 centimeters inclusive;

(c) the number of sample means falling below 172.0 centimeters.

8.23 The random variable X, representing the number of cherries in a cherry puff, has the following probability distribution:

x	4	5	6	7
$P(X = x)$	0.2	0.4	0.3	0.1

(a) Find the mean μ and the variance σ^2 of X.

(b) Find the mean $\mu_{\bar{X}}$ and the variance $\sigma^2_{\bar{X}}$ of the mean $\bar{X}$ for random samples of 36 cherry puffs.

(c) Find the probability that the average number of cherries in 36 cherry puffs will be less than 5.5.

8.24 If a certain machine makes electrical resistors having a mean resistance of 40 ohms and a standard deviation of 2 ohms, what is the probability that a random sample of 36 of these resistors will have a combined resistance of more than 1458 ohms?

8.25 The average life of a bread-making machine is 7 years, with a standard deviation of 1 year. Assuming that the lives of these machines follow approximately a normal distribution, find

(a) the probability that the mean life of a random sample of 9 such machines falls between 6.4 and 7.2 years;

(b) the value of x to the right of which 15% of the means computed from random samples of size 9 would fall.

8.26 The amount of time that a drive-through bank teller spends on a customer is a random variable with a mean $\mu = 3.2$ minutes and a standard deviation $\sigma = 1.6$ minutes. If a random sample of 64 customers is observed, find the probability that their mean time at the teller's window is

(a) at most 2.7 minutes;

(b) more than 3.5 minutes;

(c) at least 3.2 minutes but less than 3.4 minutes.

8.27 In a chemical process, the amount of a certain type of impurity in the output is difficult to control and is thus a random variable. Speculation is that the population mean amount of the impurity is 0.20 gram per gram of output. It is known that the standard deviation is 0.1 gram per gram. An experiment is conducted to gain more insight regarding the speculation that $\mu = 0.2$. The process is run on a lab scale 50 times and the sample average $\bar{x}$ turns out to be 0.23 gram per gram. Comment on the speculation that the mean amount of impurity is 0.20 gram per gram. Make use of the Central Limit Theorem in your work.

8.28 A random sample of size 25 is taken from a normal population having a mean of 80 and a standard deviation of 5. A second random sample of size 36 is taken from a different normal population having a mean of 75 and a standard deviation of 3. Find the probability that the sample mean computed from the 25 measurements will exceed the sample mean computed from the 36 measurements by at least 3.4 but less than 5.9. Assume the difference of the means to be measured to the nearest tenth.

8.29 The distribution of heights of a certain breed of terrier has a mean of 72 centimeters and a standard deviation of 10 centimeters, whereas the distribution of heights of a certain breed of poodle has a mean of 28 centimeters with a standard deviation of 5 centimeters. Assuming that the sample means can be measured to any degree of accuracy, find the probability that the sample mean for a random sample of heights of 64 terriers exceeds the sample mean for a random sample of heights of 100 poodles by at most 44.2 centimeters.

8.30 The mean score for freshmen on an aptitude test at a certain college is 540, with a standard deviation of 50. Assume the means to be measured to any degree of accuracy. What is the probability that two groups selected at random, consisting of 32 and 50 students, respectively, will differ in their mean scores by

(a) more than 20 points?

(b) an amount between 5 and 10 points?

8.31 Consider Case Study 8.2 on page 238. Suppose 18 specimens were used for each type of paint in an experiment and $\bar{x}_A - \bar{x}_B$, the actual difference in mean drying time, turned out to be 1.0.

(a) Does this seem to be a reasonable result if the

two population mean drying times truly are equal? Make use of the result in the solution to Case Study 8.2.

(b) If someone did the experiment 10,000 times under the condition that $\mu_A = \mu_B$, in how many of those 10,000 experiments would there be a difference $\bar{x}_A - \bar{x}_B$ that was as large as (or larger than) 1.0?

8.32 Two different box-filling machines are used to fill cereal boxes on an assembly line. The critical measurement influenced by these machines is the weight of the product in the boxes. Engineers are quite certain that the variance of the weight of product is $\sigma^2 = 1$ ounce. Experiments are conducted using both machines with sample sizes of 36 each. The sample averages for machines A and B are $\bar{x}_A = 4.5$ ounces and $\bar{x}_B = 4.7$ ounces. Engineers are surprised that the two sample averages for the filling machines are so different.

(a) Use the Central Limit Theorem to determine

$$P(\bar{X}_B - \bar{X}_A \geq 0.2)$$

under the condition that $\mu_A = \mu_B$.

(b) Do the aforementioned experiments seem to, in any way, strongly support a conjecture that the population means for the two machines are different? Explain using your answer in (a).

8.33 The chemical benzene is highly toxic to humans. However, it is used in the manufacture of many medicine dyes, leather, and coverings. Government regulations dictate that for any production process involving benzene, the water in the output of the process must not exceed 7950 parts per million (ppm) of benzene. For a particular process of concern, the water sample was collected by a manufacturer 25 times randomly and the sample average $\bar{x}$ was 7960 ppm. It is known from historical data that the standard deviation σ is 100 ppm.

(a) What is the probability that the sample average in this experiment would exceed the government limit if the population mean is equal to the limit? Use the Central Limit Theorem.

(b) Is an observed $\bar{x} = 7960$ in this experiment firm evidence that the population mean for the process exceeds the government limit? Answer your question by computing

$$P(\bar{X} \geq 7960 \mid \mu = 7950).$$

Assume that the distribution of benzene concentration is normal.

8.34 Two alloys A and B are being used to manufacture a certain steel product. An experiment needs to be designed to compare the two in terms of maximum load capacity in tons (the maximum weight that can be tolerated without breaking). It is known that the two standard deviations in load capacity are equal at 5 tons each. An experiment is conducted in which 30 specimens of each alloy (A and B) are tested and the results recorded as follows:

$$\bar{x}_A = 49.5, \quad \bar{x}_B = 45.5; \qquad \bar{x}_A - \bar{x}_B = 4.$$

The manufacturers of alloy A are convinced that this evidence shows conclusively that $\mu_A > \mu_B$ and strongly supports the claim that their alloy is superior. Manufacturers of alloy B claim that the experiment could easily have given $\bar{x}_A - \bar{x}_B = 4$ *even if* the two population means are equal. In other words, "the results are inconclusive!"

(a) Make an argument that manufacturers of alloy B are wrong. Do it by computing

$$P(\bar{X}_A - \bar{X}_B > 4 \mid \mu_A = \mu_B).$$

(b) Do you think these data strongly support alloy A?

8.35 Consider the situation described in Example 8.4 on page 234. Do these results prompt you to question the premise that $\mu = 800$ hours? Give a probabilistic result that indicates how *rare* an event $\bar{X} \leq 775$ is when $\mu = 800$. On the other hand, how rare would it be if μ truly were, say, 760 hours?

8.36 Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that can take on only positive values. Use the Central Limit Theorem to produce an argument that if n is sufficiently large, then $Y = X_1 X_2 \cdots X_n$ has approximately a lognormal distribution.

8.5 Sampling Distribution of S^2

In the preceding section we learned about the sampling distribution of $\bar{X}$. The Central Limit Theorem allowed us to make use of the fact that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tends toward $N(0, 1)$ as the sample size grows large. *Sampling distributions of important statistics* allow us to learn information about parameters. Usually, the parameters are the counterpart to the statistics in question. For example, if an engineer is interested in the population mean resistance of a certain type of resistor, the sampling distribution of $\bar{X}$ will be exploited once the sample information is gathered. On the other hand, if the variability in resistance is to be studied, clearly the sampling distribution of S^2 will be used in learning about the parametric counterpart, the population variance σ^2.

If a random sample of size n is drawn from a normal population with mean μ and variance σ^2, and the sample variance is computed, we obtain a value of the statistic S^2. We shall proceed to consider the distribution of the statistic $(n-1)S^2/\sigma^2$.

By the addition and subtraction of the sample mean $\bar{X}$, it is easy to see that

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2.\end{aligned}$$

Dividing each term of the equality by σ^2 and substituting $(n-1)S^2$ for $\sum_{i=1}^{n}(X_i-\bar{X})^2$, we obtain

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{(n-1)S^2}{\sigma^2}+\frac{(\bar{X}-\mu)^2}{\sigma^2/n}.$$

Now, according to Corollary 7.1 on page 222, we know that

$$\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}$$

is a chi-squared random variable with n degrees of freedom. We have a chi-squared random variable with n degrees of freedom partitioned into two components. Note that in Section 6.7 we showed that a chi-squared distribution is a special case of a gamma distribution. The second term on the right-hand side is Z^2, which is a chi-squared random variable with 1 degree of freedom, and it turns out that $(n-1)S^2/\sigma^2$ is a chi-squared random variable with $n-1$ degree of freedom. We formalize this in the following theorem.

Theorem 8.4: If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}$$

has a chi-squared distribution with $v=n-1$ degrees of freedom.

The values of the random variable χ^2 are calculated from each sample by the

formula

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}.$$

The probability that a random sample produces a χ^2 value greater than some specified value is equal to the area under the curve to the right of this value. It is customary to let χ^2_α represent the χ^2 value above which we find an area of α. This is illustrated by the shaded region in Figure 8.7.

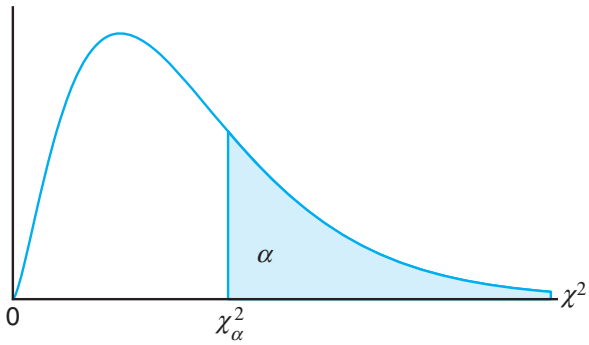

Figure 8.7: The chi-squared distribution.

Table A.5 gives values of χ^2_α for various values of α and v. The areas, α, are the column headings; the degrees of freedom, v, are given in the left column; and the table entries are the χ^2 values. Hence, the χ^2 value with 7 degrees of freedom, leaving an area of 0.05 to the right, is $\chi^2_{0.05} = 14.067$. Owing to lack of symmetry, we must also use the tables to find $\chi^2_{0.95} = 2.167$ for $v = 7$.

Exactly 95% of a chi-squared distribution lies between $\chi^2_{0.975}$ and $\chi^2_{0.025}$. A χ^2 value falling to the right of $\chi^2_{0.025}$ is not likely to occur unless our assumed value of σ^2 is too small. Similarly, a χ^2 value falling to the left of $\chi^2_{0.975}$ is unlikely unless our assumed value of σ^2 is too large. In other words, it is possible to have a χ^2 value to the left of $\chi^2_{0.975}$ or to the right of $\chi^2_{0.025}$ when σ^2 is correct, but if this should occur, it is more probable that the assumed value of σ^2 is in error.

Example 8.7: A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.

***Solution*:** We first find the sample variance using Theorem 8.1,

$$s^2 = \frac{(5)(48.26) - (15)^2}{(5)(4)} = 0.815.$$

Then

$$\chi^2 = \frac{(4)(0.815)}{1} = 3.26$$

is a value from a chi-squared distribution with 4 degrees of freedom. Since 95% of the χ^2 values with 4 degrees of freedom fall between 0.484 and 11.143, the computed value with $\sigma^2 = 1$ is reasonable, and therefore the manufacturer has no reason to suspect that the standard deviation is other than 1 year.

Degrees of Freedom as a Measure of Sample Information

Recall from Corollary 7.1 in Section 7.3 that

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$$

has a χ^2-distribution with n *degrees of freedom*. Note also Theorem 8.4, which indicates that the random variable

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

has a χ^2-distribution with $n-1$ *degrees of freedom*. The reader may also recall that the term *degrees of freedom*, used in this identical context, is discussed in Chapter 1.

As we indicated earlier, the proof of Theorem 8.4 will not be given. However, the reader can view Theorem 8.4 as indicating that when μ is not known and one considers the distribution of

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2},$$

there is **1 less degree of freedom**, or a degree of freedom is lost in the estimation of μ (i.e., when μ is replaced by $\bar{x}$). In other words, there are n degrees of freedom, or independent *pieces of information*, in the random sample from the normal distribution. When the data (the values in the sample) are used to compute the mean, there is 1 less degree of freedom in the information used to estimate σ^2.