CHAPTER 5

PROBABILISTIC FEATURES OF THE DISTRIBUTIONS OF CERTAIN SAMPLE STATISTICS

CHAPTER OVERVIEW

This chapter ties together the foundations of applied statistics: descriptive measures, basic probability, and inferential procedures. This chapter also includes a discussion of one of the most important theorems in statistics, the central limit theorem. Students may find it helpful to revisit this chapter from time to time as they study the remaining chapters of the book.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. be able to construct a sampling distribution of a statistic.
2. understand how to use a sampling distribution to calculate basic probabilities.
3. understand the central limit theorem and when to apply it.
4. understand the basic concepts of sampling with replacement and without replacement.

5.1 INTRODUCTION

Before we examine the subject matter of this chapter, let us review the high points of what we have covered thus far. Chapter 1 introduces some basic and useful statistical vocabulary and discusses the basic concepts of data collection. In Chapter 2, the organization and summarization of data are emphasized. It is here that we encounter the concepts of central tendency and dispersion and learn how to compute their descriptive measures. In Chapter 3, we are introduced to the fundamental ideas of probability, and in Chapter 4 we consider the concept of a probability distribution. These concepts are fundamental to an understanding of statistical inference, the topic that comprises the major portion of this book.

This chapter serves as a bridge between the preceding material, which is essentially descriptive in nature, and most of the remaining topics, which have been selected from the area of statistical inference.

5.2 SAMPLING DISTRIBUTIONS

The topic of this chapter is *sampling distributions*. The importance of a clear understanding of sampling distributions cannot be overemphasized, as this concept is the very key to the understanding of statistical inference. Sampling distributions serve two purposes: (1) they allow us to answer probability questions about sample statistics, and (2) they provide the necessary theory for making statistical inference procedures valid. In this chapter we use sampling distributions to answer probability questions about sample statistics. We recall from Chapter 2 that a sample statistic is a descriptive measure, such as the mean, median, variance, or standard deviation, that is computed from the data of a sample. In the chapters that follow, we will see how sampling distributions make statistical inferences valid.

We begin with the following definition.

DEFINITION

The distribution of all possible values that can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population, is called the *sampling distribution* of that statistic.

Sampling Distributions: Construction Sampling distributions may be constructed empirically when sampling from a discrete, finite population. To construct a sampling distribution we proceed as follows:

1. From a finite population of size N, randomly draw all possible samples of size n.
2. Compute the statistic of interest for each sample.
3. List in one column the different distinct observed values of the statistic, and in another column list the corresponding frequency of occurrence of each distinct observed value of the statistic.

The actual construction of a sampling distribution is a formidable undertaking if the population is of any appreciable size and is an impossible task if the population is infinite. In such cases, sampling distributions may be approximated by taking a large number of samples of a given size.

Sampling Distributions: Important Characteristics We usually are interested in knowing three things about a given sampling distribution: its *mean*, its *variance*, and its *functional form* (how it looks when graphed).

We can recognize the difficulty of constructing a sampling distribution according to the steps given above when the population is large. We also run into a problem when considering the construction of a sampling distribution when the population is infinite. The best we can do experimentally in this case is to approximate the sampling distribution of a statistic.

Both these problems may be obviated by means of mathematics. Although the procedures involved are not compatible with the mathematical level of this text, sampling distributions can be derived mathematically. The interested reader can consult one of many mathematical statistics textbooks, for example, Larsen and Marx (1) or Rice (2).

In the sections that follow, some of the more frequently encountered sampling distributions are discussed.

5.3 DISTRIBUTION OF THE SAMPLE MEAN

An important sampling distribution is the distribution of the sample mean. Let us see how we might construct the sampling distribution by following the steps outlined in the previous section.

EXAMPLE 5.3.1

Suppose we have a population of size $N = 5$, consisting of the ages of five children who are outpatients in a community mental health center. The ages are as follows: $x_1 = 6, x_2 = 8, x_3 = 10, x_4 = 12$, and $x_5 = 14$. The mean, μ, of this population is equal to $\Sigma x_i/N = 10$ and the variance is

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} = \frac{40}{5} = 8$$

Let us compute another measure of dispersion and designate it by capital S as follows:

$$S^2 = \frac{\Sigma(x_i - \mu)^2}{N - 1} = \frac{40}{4} = 10$$

We will refer to this quantity again in the next chapter. We wish to construct the sampling distribution of the sample mean, $\bar{x}$, based on samples of size $n = 2$ drawn from this population.

Solution: Let us draw all possible samples of size $n = 2$ from this population. These samples, along with their means, are shown in Table 5.3.1.

TABLE 5.3.1 All Possible Samples of Size $n = 2$ from a Population of Size $N = 5$. Samples Above or Below the Principal Diagonal Result When Sampling Is Without Replacement. Sample Means Are in Parentheses

		Second Draw				
		6	8	10	12	14
First Draw	6	6, 6 (6)	6, 8 (7)	6, 10 (8)	6, 12 (9)	6, 14 (10)
	8	8, 6 (7)	8, 8 (8)	8, 10 (9)	8, 12 (10)	8, 14 (11)
	10	10, 6 (8)	10, 8 (9)	10, 10 (10)	10, 12 (11)	10, 14 (12)
	12	12, 6 (9)	12, 8 (10)	12, 10 (11)	12, 12 (12)	12, 14 (13)
	14	14, 6 (10)	14, 8 (11)	14, 10 (12)	14, 12 (13)	14, 14 (14)

TABLE 5.3.2 Sampling Distribution of $\bar{x}$ Computed from Samples in Table 5.3.1

$\bar{x}$	Frequency	Relative Frequency
6	1	1/25
7	2	2/25
8	3	3/25
9	4	4/25
10	5	5/25
11	4	4/25
12	3	3/25
13	2	2/25
14	1	1/25
Total	25	25/25

We see in this example that, when sampling is with replacement, there are 25 possible samples. In general, when sampling is with replacement, the number of possible samples is equal to N^n.

We may construct the sampling distribution of $\bar{x}$ by listing the different values of $\bar{x}$ in one column and their frequency of occurrence in another, as in Table 5.3.2. ■

We see that the data of Table 5.3.2 satisfy the requirements for a probability distribution. The individual probabilities are all greater than 0, and their sum is equal to 1.

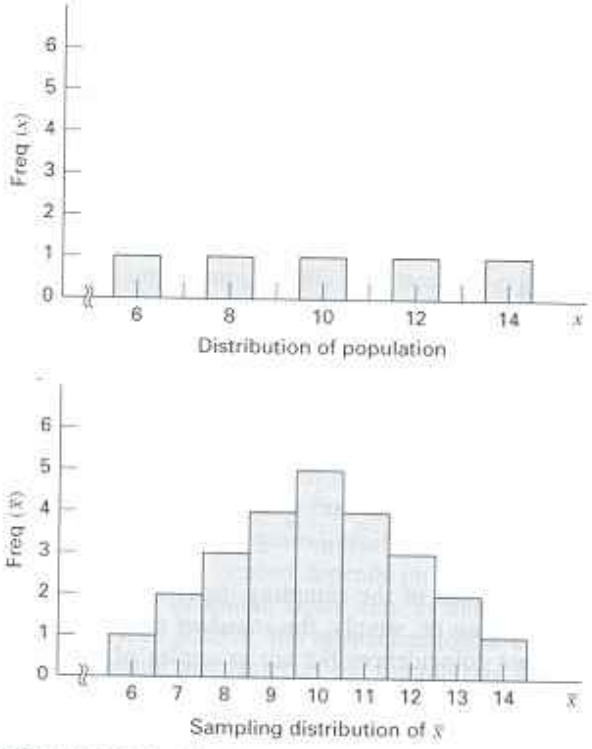

FIGURE 5.3.1 Distribution of population and sampling distribution of $\bar{x}$.

It was stated earlier that we are usually interested in the functional form of a sampling distribution, its mean, and its variance. We now consider these characteristics for the sampling distribution of the sample mean, $\bar{x}$.

Sampling Distribution of $\bar{x}$: Functional Form Let us look at the distribution of $\bar{x}$ plotted as a histogram, along with the distribution of the population, both of which are shown in Figure 5.3.1. We note the radical difference in appearance between the histogram of the population and the histogram of the sampling distribution of $\bar{x}$. Whereas the former is uniformly distributed, the latter gradually rises to a peak and then drops off with perfect symmetry.

Sampling Distribution of $\bar{x}$: Mean Now let us compute the mean, which we will call $\mu_{\bar{x}}$, of our sampling distribution. To do this we add the 25 sample means and divide by 25. Thus,

$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{N''} = \frac{6 + 7 + 7 + 8 + \cdots + 14}{25} = \frac{250}{25} = 10$$

We note with interest that the mean of the sampling distribution of $\bar{x}$ has the same value as the mean of the original population.

Sampling Distribution of $\bar{x}$: Variance Finally, we may compute the variance of $\bar{x}$, which we call $\sigma_{\bar{x}}^2$ as follows:

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \frac{\sum(\bar{x}_i - \mu_{\bar{x}})^2}{N^n} \\ &= \frac{(6-10)^2 + (7-10)^2 + (7-10)^2 + \cdots + (14-10)^2}{25} \\ &= \frac{100}{25} = 4\end{aligned}$$

We note that the variance of the sampling distribution is not equal to the population variance. It is of interest to observe, however, that the variance of the sampling distribution is equal to the population variance divided by the size of the sample used to obtain the sampling distribution. That is,

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{8}{2} = 4$$

The square root of the variance of the sampling distribution, $\sqrt{\sigma_{\bar{x}}^2} = \sigma/\sqrt{n}$ is called the *standard error of the mean* or, simply, the *standard error.*

These results are not coincidences but are examples of the characteristics of sampling distributions in general, when sampling is with replacement or when sampling is from an infinite population. To generalize, we distinguish between two situations: sampling from a normally distributed population and sampling from a nonnormally distributed population.

Sampling Distribution of $\bar{x}$: Sampling from Normally Distributed Populations When sampling is from a normally distributed population, the distribution of the sample mean will possess the following properties:

1. The distribution of $\bar{x}$ will be normal.
2. The mean, $\mu_{\bar{x}}$, of the distribution of $\bar{x}$ will be equal to the mean of the population from which the samples were drawn.
3. The variance, $\sigma_{\bar{x}}^2$ of the distribution of $\bar{x}$ will be equal to the variance of the population divided by the sample size.

Sampling from Nonnormally Distributed Populations For the case where sampling is from a nonnormally distributed population, we refer to an important mathematical theorem known as the *central limit theorem.* The importance of this theorem in statistical inference may be summarized in the following statement.

The Central Limit Theorem

> *Given a population of any nonnormal functional form with a mean μ and finite variance σ^2, the sampling distribution of $\bar{x}$, computed from samples of size n from this population, will have mean μ and variance σ^2/n and will be approximately normally distributed when the sample size is large.*

A mathematical formulation of the central limit theorem is that the distribution of

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

approaches a normal distribution with mean 0 and variance 1 as $n \to \infty$. Note that the central limit theorem allows us to sample from nonnormally distributed populations with a guarantee of approximately the same results as would be obtained if the populations were normally distributed provided that we take a large sample.

The importance of this will become evident later when we learn that a normally distributed sampling distribution is a powerful tool in statistical inference. In the case of the sample mean, we are assured of at least an approximately normally distributed sampling distribution under three conditions: (1) when sampling is from a normally distributed population; (2) when sampling is from a nonnormally distributed population and our sample is large; and (3) when sampling is from a population whose functional form is unknown to us as long as our sample size is large.

The logical question that arises at this point is, How large does the sample have to be in order for the central limit theorem to apply? There is no one answer, since the size of the sample needed depends on the extent of nonnormality present in the population. One rule of thumb states that, in most practical situations, a sample of size 30 is satisfactory. In general, the approximation to normality of the sampling distribution of $\bar{x}$ becomes better and better as the sample size increases.

Sampling Without Replacement The foregoing results have been given on the assumption that sampling is either with replacement or that the samples are drawn from infinite populations. In general, we do not sample with replacement, and in most practical situations it is necessary to sample from a finite population; hence, we need to become familiar with the behavior of the sampling distribution of the sample mean under these conditions. Before making any general statements, let us again look at the data in Table 5.3.1. The sample means that result when sampling is without replacement are those above the principal diagonal, which are the same as those below the principal diagonal, if we ignore the order in which the observations were drawn. We see that there are 10 possible samples. In general, when drawing samples of size n from a finite population of size N without replacement, and ignoring the order in which the sample values are drawn, the number of possible samples is given by the combination of N things taken n at a time. In our present example we have

$${}_NC_n = \frac{N!}{n!(N-n)!} = \frac{5!}{2!3!} = \frac{5 \cdot 4 \cdot 3!}{2!3!} = 10 \text{ possible samples.}$$

The mean of the 10 sample means is

$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{{}_NC_n} = \frac{7 + 8 + 9 + \cdots + 13}{10} = \frac{100}{10} = 10$$

We see that once again the mean of the sampling distribution is equal to the population mean.

The variance of this sampling distribution is found to be

$$\sigma_{\bar{x}}^2 = \frac{\sum(\bar{x}_i - \mu_{\bar{x}})^2}{{}_NC_n} = \frac{30}{10} = 3$$

and we note that this time the variance of the sampling distribution is not equal to the population variance divided by the sample size, since $\sigma_{\bar{x}}^2 = 3 \neq 8/2 = 4$. There is, however, an interesting relationship that we discover by multiplying σ^2/n by $(N - n)/(N - 1)$. That is,

$$\frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1} = \frac{8}{2} \cdot \frac{5 - 2}{4} = 3$$

This result tells us that if we multiply the variance of the sampling distribution that would be obtained if sampling were with replacement, by the factor $(N - n)/(N - 1)$, we obtain the value of the variance of the sampling distribution that results when sampling is without replacement. We may generalize these results with the following statement.

When sampling is without replacement from a finite population, the sampling distribution of $\bar{x}$ will have mean μ and variance

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}$$

If the sample size is large, the central limit theorem applies and the sampling distribution of $\bar{x}$ will be approximately normally distributed.

The Finite Population Correction The factor $(N - n)/(N - 1)$ is called the finite population correction and can be ignored when the sample size is small in comparison with the population size. When the population is much larger than the sample, the difference between σ^2/n and $(\sigma^2/n)[(N - n)/(N - 1)]$ will be negligible. Imagine a population of size 10,000 and a sample from this population of size 25; the finite population correction would be equal to $(10{,}000 - 25)/(9999) = .9976$. To multiply σ^2/n by .9976 is almost equivalent to multiplying it by 1. Most practicing statisticians do not use the finite population correction unless the sample is more than 5 percent of the size of the population. That is, the finite population correction is usually ignored when $n/N \leq .05$.

The Sampling Distribution of $\bar{x}$: A Summary Let us summarize the characteristics of the sampling distribution of $\bar{x}$ under two conditions.

1. Sampling is from a normally distributed population with a known population variance:
   (a) $\mu_{\bar{x}} = \mu$
   (b) $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
   (c) The sampling distribution of $\bar{x}$ is normal.

2. Sampling is from a nonnormally distributed population with a known population variance:
   (a) $\mu_{\bar{x}} = \mu$
   (b) $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, when $n/N \leq .05$
   $\sigma_{\bar{x}} = (\sigma/\sqrt{n})\sqrt{\dfrac{N-n}{N-1}}$, otherwise
   (c) The sampling distribution of $\bar{x}$ is approximately normal.

Applications As we will see in succeeding chapters, knowledge and understanding of sampling distributions will be necessary for understanding the concepts of statistical inference. The simplest application of our knowledge of the sampling distribution of the sample mean is in computing the probability of obtaining a sample with a mean of some specified magnitude. Let us illustrate with some examples.

EXAMPLE 5.3.2

Suppose it is known that in a certain large human population cranial length is approximately normally distributed with a mean of 185.6 mm and a standard deviation of 12.7 mm. What is the probability that a random sample of size 10 from this population will have a mean greater than 190?

Solution: We know that the single sample under consideration is one of all possible samples of size 10 that can be drawn from the population, so that the mean that it yields is one of the $\bar{x}$'s constituting the sampling distribution of $\bar{x}$ that, theoretically, could be derived from this population.

When we say that the population is approximately normally distributed, we assume that the sampling distribution of $\bar{x}$ will be, for all practical purposes, normally distributed. We also know that the mean and standard deviation of the sampling distribution are equal to 185.6 and $\sqrt{(12.7)^2/10} = 12.7/\sqrt{10} = 4.0161$, respectively. We assume that the population is large relative to the sample so that the finite population correction can be ignored.

We learn in Chapter 4 that whenever we have a random variable that is normally distributed, we may very easily transform it to the standard normal distribution. Our random variable now is $\bar{x}$, the mean of its distribution is $\mu_{\bar{x}}$, and its standard deviation is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. By appropriately modifying the formula given previously, we arrive at the following formula for transforming the normal distribution of $\bar{x}$ to the standard normal distribution:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma/\sqrt{n}} \tag{5.3.1}$$

■

The probability that answers our question is represented by the area to the right of $\bar{x} = 190$ under the curve of the sampling distribution. This area is equal to the area to the right of

$$z = \frac{190 - 185.6}{4.0161} = \frac{4.4}{4.0161} = 1.10$$

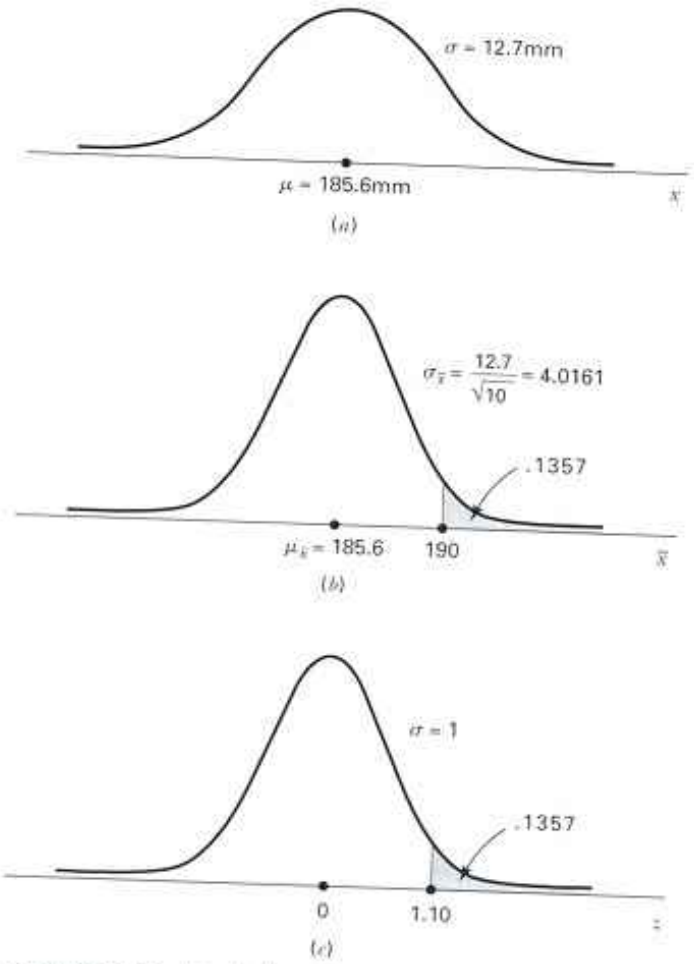

FIGURE 5.3.2 Population distribution, sampling distribution, and standard normal distribution, Example 5.3.2: (*a*) population distribution; (*b*) sampling distribution of $\bar{x}$ for samples of size 10; (*c*) standard normal distribution.

By consulting the standard normal table, we find that the area to the right of 1.10 is .1357; hence, we say that the probability is .1357 that a sample of size 10 will have a mean greater than 190.

Figure 5.3.2 shows the relationship between the original population, the sampling distribution of $\bar{x}$ and the standard normal distribution.

EXAMPLE 5.3.3

If the mean and standard deviation of serum iron values for healthy men are 120 and 15 micrograms per 100 ml, respectively, what is the probability that a random sample of 50 normal men will yield a mean between 115 and 125 micrograms per 100 ml?

Solution: The functional form of the population of serum iron values is not specified, but since we have a sample size greater than 30, we make use of the

central limit theorem and transform the resulting approximately normal sampling distribution of $\bar{x}$ (which has a mean of 120 and a standard deviation of $15/\sqrt{50} = 2.1213$) to the standard normal. The probability we seek is

$$\begin{aligned} P(115 \le \bar{x} \le 125) &= P\left(\frac{115 - 120}{2.12} \le z \le \frac{125 - 120}{2.12}\right) \\ &= P(-2.36 \le z \le 2.36) \\ &= .9909 - .0091 \\ &= .9818 \end{aligned}$$

■

EXERCISES

5.3.1 The National Health and Nutrition Examination Survey of 1988–1994 (NHANES III, A-1) estimated the mean serum cholesterol level for U.S. females aged 20–74 years to be 204 mg/dl. The estimate of the standard deviation was approximately 44. Using these estimates as the mean μ and standard deviation σ for the U.S. population, consider the sampling distribution of the sample mean based on samples of size 50 drawn from women in this age group. What is the mean of the sampling distribution? The standard error?

5.3.2 The study cited in Exercise 5.3.1 reported an estimated mean serum cholesterol level of 183 for women aged 20–29 years. The estimated standard deviation was approximately 37. Use these estimates as the mean μ and standard deviation σ for the U.S. population. If a simple random sample of size 60 is drawn from this population, find the probability that the sample mean serum cholesterol level will be:

(a) Between 170 and 195 (b) Below 175

(c) Greater than 190

5.3.3 If the uric acid values in normal adult males are approximately normally distributed with a mean and standard deviation of 5.7 and 1 mg percent, respectively, find the probability that a sample of size 9 will yield a mean:

(a) Greater than 6 (b) Between 5 and 6

(c) Less than 5.2

5.3.4 Wright et al. (A-2) used the 1999–2000 National Health and Nutrition Examination Survey (NHANES) to estimate dietary intake of 10 key nutrients. One of those nutrients was calcium (mg). They found in all adults 60 years or older a mean daily calcium intake of 721 mg with a standard deviation of 454. Using these values for the mean and standard deviation for the U.S. population, find the probability that a random sample of size 50 will have a mean:

(a) Greater than 800 mg (b) Less than 700 mg

(c) Between 700 and 850 mg

5.3.5 In the study cited in Exercise 5.3.4, researchers found the mean sodium intake in men and women 60 years or older to be 2940 mg with a standard deviation of 1476 mg. Use these values for the mean and standard deviation of the U.S. population and find the probability that a random sample of 75 people from the population will have a mean:

(a) Less than 2450 mg (b) Over 3100 mg

(c) Between 2500 and 3300 mg (d) Between 2500 and 2900 mg

5.3.6 Given a normally distributed population with a mean of 100 and a standard deviation of 20, find the following probabilities based on a sample of size 16:

(a) $P(\bar{x} \geq 100)$ (b) $P(\bar{x} \leq 110)$

(c) $P(96 \leq \bar{x} \leq 108)$

5.3.7 Given $\mu = 50$, $\sigma = 16$, and $n = 64$, find:

(a) $P(45 \leq \bar{x} \leq 55)$ (b) $P(\bar{x} > 53)$

(c) $P(\bar{x} < 47)$ (d) $P(49 \leq \bar{x} \leq 56)$

5.3.8 Suppose a population consists of the following values: 1, 3, 5, 7, 9. Construct the sampling distribution of $\bar{x}$ based on samples of size 2 selected without replacement. Find the mean and variance of the sampling distribution.

5.3.9 Use the data of Example 5.3.1 to construct the sampling distribution of $\bar{x}$ based on samples of size 3 selected without replacement. Find the mean and variance of the sampling distribution.

5.3.10 Use the data cited in Exercise 5.3.1. Imagine we take samples of size 5, 25, 50, 100, and 500 from the women in this age group.

(a) Calculate the standard error for each of these sampling scenarios.

(b) Discuss how sample size affects the standard error estimates calculated in part (a) and the potential implications this may have in statistical practice.

5.4 DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS

Frequently the interest in an investigation is focused on two populations. Specifically, an investigator may wish to know something about the difference between two population means. In one investigation, for example, a researcher may wish to know if it is reasonable to conclude that two population means are different. In another situation, the researcher may desire knowledge about the magnitude of the difference between two population means. A medical research team, for example, may want to know whether or not the mean serum cholesterol level is higher in a population of sedentary office workers than in a population of laborers. If the researchers are able to conclude that the population means are different, they may wish to know by how much they differ. A knowledge of the sampling distribution of the difference between two means is useful in investigations of this type.

Sampling from Normally Distributed Populations The following example illustrates the construction of and the characteristics of the sampling distribution of the difference between sample means when sampling is from two normally distributed populations.

EXAMPLE 5.4.1

Suppose we have two populations of individuals—one population (population 1) has experienced some condition thought to be associated with mental retardation, and the other population (population 2) has not experienced the condition. The distribution of

intelligence scores in each of the two populations is believed to be approximately normally distributed with a standard deviation of 20.

Suppose, further, that we take a sample of 15 individuals from each population and compute for each sample the mean intelligence score with the following results: $\bar{x}_1 = 92$ and $\bar{x}_2 = 105$. If there is no difference between the two populations, with respect to their true mean intelligence scores, what is the probability of observing a difference this large or larger $(\bar{x}_1 - \bar{x}_2)$ between sample means?

Solution: To answer this question we need to know the nature of the sampling distribution of the relevant statistic, the *difference between two sample means*, $\bar{x}_1 - \bar{x}_2$. Notice that we seek a probability associated with the difference between two sample means rather than a single mean. ■

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$: Construction Although, in practice, we would not attempt to construct the desired sampling distribution, we can conceptualize the manner in which it could be done when sampling is from finite populations. We would begin by selecting from population 1 all possible samples of size 15 and computing the mean for each sample. We know that there would be ${}_{N_1}C_{n_1}$ such samples where N_1 is the population size and $n_1 = 15$. Similarly, we would select all possible samples of size 15 from population 2 and compute the mean for each of these samples. We would then take all possible pairs of sample means, one from population 1 and one from population 2, and take the difference. Table 5.4.1 shows the results of following this procedure. Note that the 1's and 2's in the last line of this table are not exponents, but indicators of population 1 and 2, respectively.

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$: Characteristics It is the distribution of the differences between sample means that we seek. If we plotted the sample differences against their frequency of occurrence, we would obtain a normal distribution with a mean equal to $\mu_1 - \mu_2$, the difference between the two population means, and a variance equal to $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$. That is, the standard error of the difference between

TABLE 5.4.1 Working Table for Constructing the Distribution of the Difference Between Two Sample Means

Samples from Population 1	Samples from Population 2	Sample Means Population 1	Sample Means Population 2	All Possible Differences Between Means
n_{11}	n_{12}	$\bar{x}_{11}$	$\bar{x}_{12}$	$\bar{x}_{11} - \bar{x}_{12}$
n_{21}	n_{22}	$\bar{x}_{21}$	$\bar{x}_{22}$	$\bar{x}_{11} - \bar{x}_{22}$
n_{31}	n_{32}	$\bar{x}_{31}$	$\bar{x}_{32}$	$\bar{x}_{11} - \bar{x}_{32}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$n_{N_1C_{n_1}}1$	$n_{N_2C_{n_2}}2$	$\bar{x}_{N_1C_{n_1}}1$	$\bar{x}_{N_2C_{n_2}}2$	$\bar{x}_{N_1C_{n_1}}1 - \bar{x}_{N_2C_{n_2}}2$

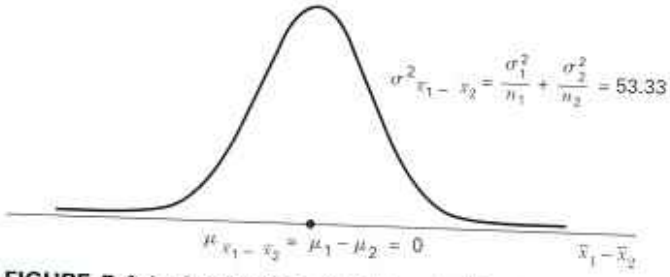

FIGURE 5.4.1 Graph of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ when there is no difference between population means, Example 5.4.1.

sample means would be equal to $\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}$. It should be noted that these properties convey two important points. First, the means of two distributions can be subtracted from one another, or summed together, using standard arithmetic operations. Second, since the overall variance of the sampling distribution will be affected by both contributing distributions, the variances will always be summed even if we are interested in the difference of the means. This last fact assumes that the two distributions are independent of one another.

For our present example we would have a normal distribution with a mean of 0 (if there is no difference between the two population means) and a variance of $[(20)^2/15] + [(20)^2/15] = 53.3333$. The graph of the sampling distribution is shown in Figure 5.4.1.

Converting to *z* We know that the normal distribution described in Example 5.4.1 can be transformed to the standard normal distribution by means of a modification of a previously learned formula. The new formula is as follows:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{5.4.1}$$

The area under the curve of $\bar{x}_1 - \bar{x}_2$ corresponding to the probability we seek is the area to the left of $\bar{x}_1 - \bar{x}_2 = 92 - 105 = -13$. The z value corresponding to -13, assuming that there is no difference between population means, is

$$z = \frac{-13 - 0}{\sqrt{\dfrac{(20)^2}{15} + \dfrac{(20)^2}{15}}} = \frac{-13}{\sqrt{53.3}} = \frac{-13}{7.3} = -1.78$$

By consulting Table D, we find that the area under the standard normal curve to the left of -1.78 is equal to .0375. In answer to our original question, we say that if there is no

difference between population means, the probability of obtaining a difference between sample means as large as or larger than 13 is .0375.

Sampling from Normal Populations The procedure we have just followed is valid even when the sample sizes, n_1 and n_2, are different and when the population variances, σ_1^2 and σ_2^2 have different values. The theoretical results on which this procedure is based may be summarized as follows.

Given two normally distributed populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, the sampling distribution of the difference, $\bar{x}_1 - \bar{x}_2$, between the means of independent samples of size n_1 and n_2 drawn from these populations is normally distributed with mean $\mu_1 - \mu_2$ and variance $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$.

Sampling from Nonnormal Populations Many times a researcher is faced with one or the other of the following problems: the necessity of (1) sampling from nonnormally distributed populations, or (2) sampling from populations whose functional forms are not known. A solution to these problems is to take large samples, since when the sample sizes are large the central limit theorem applies and the distribution of the difference between two sample means is at least approximately normally distributed with a mean equal to $\mu_1 - \mu_2$ and a variance of $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$. To find probabilities associated with specific values of the statistic, then, our procedure would be the same as that given when sampling is from normally distributed populations.

EXAMPLE 5.4.2

Suppose it has been established that for a certain type of client the average length of a home visit by a public health nurse is 45 minutes with a standard deviation of 15 minutes, and that for a second type of client the average home visit is 30 minutes long with a standard deviation of 20 minutes. If a nurse randomly visits 35 clients from the first and 40 from the second population, what is the probability that the average length of home visit will differ between the two groups by 20 or more minutes?

Solution: No mention is made of the functional form of the two populations, so let us assume that this characteristic is unknown, or that the populations are not normally distributed. Since the sample sizes are large (greater than 30) in both cases, we draw on the results of the central limit theorem to answer the question posed. We know that the difference between sample means is at least approximately normally distributed with the following mean and variance:

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 = 45 - 30 = 15$$

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(15)^2}{35} + \frac{(20)^2}{40} = 16.4286$$

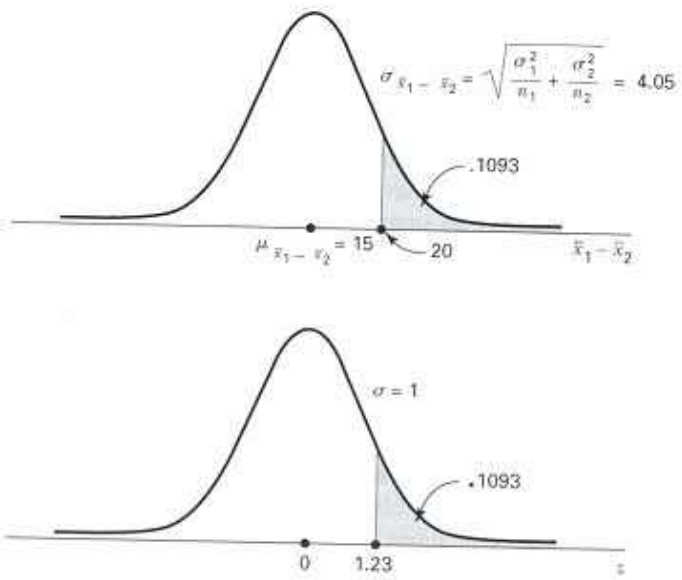

FIGURE 5.4.2 Sampling distribution of $\bar{x}_1 - \bar{x}_2$ and the corresponding standard normal distribution, home visit example.

The area under the curve of $\bar{x}_1 - \bar{x}_2$ that we seek is that area to the right of 20. The corresponding value of z in the standard normal is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{20 - 15}{\sqrt{16.4286}} = \frac{5}{4.0532} = 1.23$$

In Table D we find that the area to the right of $z = 1.23$ is $1 - .8907 = .1093$. We say, then, that the probability of the nurse's random visits resulting in a difference between the two means as great as or greater than 20 minutes is .1093. The curve of $\bar{x}_1 - \bar{x}_2$ and the corresponding standard normal curve are shown in Figure 5.4.2. ■

EXERCISES

5.4.1 The study cited in Exercises 5.3.1 and 5.3.2 gives the following data on serum cholesterol levels in U.S. females:

Population	Age	Mean	Standard Deviation
A	20–29	183	37.2
B	30–39	189	34.7

Use these estimates as the mean μ and standard deviation σ for the respective U.S. populations. Suppose we select a simple random sample of size 50 independently from each population. What is the probability that the difference between sample means $\bar{x}_B - \bar{x}_A$ will be more than 8?

5.4.2 In the study cited in Exercises 5.3.4 and 5.3.5, the calcium levels in men and women ages 60 years or older are summarized in the following table:

	Mean	Standard Deviation
Men	797	482
Women	660	414

Use these estimates as the mean μ and standard deviation σ for the U.S. populations for these age groups. If we take a random sample of 40 men and 35 women, what is the probability of obtaining a difference between sample means of 100 mg or more?

5.4.3 Given two normally distributed populations with equal means and variances of $\sigma_1^2 = 100$ and $\sigma_2^2 = 80$, what is the probability that samples of size $n_1 = 25$ and $n_2 = 16$ will yield a value of $\bar{x}_1 - \bar{x}_2$ greater than or equal to 8?

5.4.4 Given two normally distributed populations with equal means and variances of $\sigma_1^2 = 240$ and $\sigma_2^2 = 350$, what is the probability that samples of size $n_1 = 40$ and $n_2 = 35$ will yield a value of $\bar{x}_1 - \bar{x}_2$ as large as or larger than 12?

5.4.5 For a population of 17-year-old boys and 17-year-old girls, the means and standard deviations, respectively, of their subscapular skinfold thickness values are as follows: boys, 9.7 and 6.0; girls, 15.6 and 9.5. Simple random samples of 40 boys and 35 girls are selected from the populations. What is the probability that the difference between sample means $\bar{x}_{\text{girls}} - \bar{x}_{\text{boys}}$ will be greater than 10?

5.5 DISTRIBUTION OF THE SAMPLE PROPORTION

In the previous sections we have dealt with the sampling distributions of statistics computed from measured variables. We are frequently interested, however, in the sampling distribution of a statistic, such as a sample proportion, that results from counts or frequency data.

EXAMPLE 5.5.1

Results (A-3) from the 1999–2000 National Health and Nutrition Examination Survey (NHANES), show that 31 percent of U.S. adults ages 20–74 are obese (obese as defined with body mass index greater than or equal to 30.0). We designate this population proportion as $p = .31$. If we randomly select 150 individuals from this population, what is the probability that the proportion in the sample who are obese will be as great as .40?

Solution: To answer this question, we need to know the properties of the sampling distribution of the sample proportion. We will designate the sample proportion by the symbol $\hat{p}$.

You will recognize the similarity between this example and those presented in Section 4.3, which dealt with the binomial distribution. The variable obesity is a *dichotomous variable*, since an individual can be classified into one or the other of two mutually exclusive categories obese or not obese. In Section 4.3, we were given similar information and were asked to find the number with the characteristic of interest, whereas here we are seeking the proportion in the sample possessing the characteristic of interest. We could with a sufficiently large table of binomial probabilities, such as Table B, determine the probability associated with the number corresponding to the proportion of interest. As we will see, this will not be necessary, since there is available an alternative procedure, when sample sizes are large, that is generally more convenient. ■

Sampling Distribution of $\hat{p}$: Construction The sampling distribution of a sample proportion would be constructed experimentally in exactly the same manner as was suggested in the case of the arithmetic mean and the difference between two means. From the population, which we assume to be finite, we would take all possible samples of a given size and for each sample compute the sample proportion, $\hat{p}$. We would then prepare a frequency distribution of $\hat{p}$ by listing the different distinct values of $\hat{p}$ along with their frequencies of occurrence. This frequency distribution (as well as the corresponding relative frequency distribution) would constitute the sampling distribution of $\hat{p}$.

Sampling Distribution of $\hat{p}$: Characteristics When the sample size is large, the distribution of sample proportions is approximately normally distributed by virtue of the central limit theorem. The mean of the distribution, $\mu_{\hat{p}}$, that is, the average of all the possible sample proportions, will be equal to the true population proportion, p, and the variance of the distribution, $\sigma^2_{\hat{p}}$, will be equal to $p(1-p)/n$ or pq/n, where $q = 1 - p$. To answer probability questions about p, then, we use the following formula:

$$z = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \tag{5.5.1}$$

The question that now arises is, How large does the sample size have to be for the use of the normal approximation to be valid? A widely used criterion is that both np and $n(1-p)$ must be greater than 5, and we will abide by that rule in this text.

We are now in a position to answer the question regarding obesity in the sample of 150 individuals from a population in which 31 percent are obese. Since both np and $n(1-p)$ are greater than $5(150 \times .31 = 46.5$ and $150 \times .69 = 103.5)$, we can say that, in this case, $\hat{p}$ is approximately normally distributed with a mean $\mu_{\hat{p}} = p = .31$ and $\sigma^2_{\hat{p}} = p(1-p)/n = (.31)(.69)/150 = .001426$. The probability we seek is the area

under the curve of $\hat{p}$ that is to the right of .40. This area is equal to the area under the standard normal curve to the right of

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.40 - .31}{\sqrt{.001426}} = 2.38$$

The transformation to the standard normal distribution has been accomplished in the usual manner: z is found by dividing the difference between a value of a statistic and its mean by the standard error of the statistic. Using Table D we find that the area to the right of $z = 2.38$ is $1 - .9913 = .0087$. We may say, then, that the probability of observing $\hat{p} \geq .40$ in a random sample of size $n = 150$ from a population in which $p = .31$ is .0087. If we should, in fact, draw such a sample, most people would consider it a rare event.

Correction for Continuity The normal approximation may be improved by the *correction for continuity*, a device that makes an adjustment for the fact that a discrete distribution is being approximated by a continuous distribution. Suppose we let $x = n\hat{p}$, the number in the sample with the characteristic of interest when the proportion is $\hat{p}$. To apply the correction for continuity, we compute

$$z_c = \frac{\frac{x + .5}{n} - p}{\sqrt{pq/n}}, \qquad \text{for } x < np \tag{5.5.2}$$

or

$$z_c = \frac{\frac{x - .5}{n} - p}{\sqrt{pq/n}}, \qquad \text{for } x > np \tag{5.5.3}$$

where $q = 1 - p$. The correction for continuity will not make a great deal of difference when n is large. In the above example $n\hat{p} = 150(.4) = 60$, and

$$z_c = \frac{\frac{60 - .5}{150} - .31}{\sqrt{(.31)(.69)/150}} = 2.30$$

and $P(\hat{p} \geq .40) = 1 - .9893 = .0107$, a result not greatly different from that obtained without the correction for continuity. This adjustment is not often done by hand, since most statistical computer programs automatically apply the appropriate continuity correction when necessary.

EXAMPLE 5.5.2

Blanche Mikhail (A-4) studied the use of prenatal care among low-income African-American women. She found that only 51 percent of these women had adequate prenatal care. Let us assume that for a population of similar low-income African-American women, 51 percent had adequate prenatal care. If 200 women from this population are drawn at random, what is the probability that less than 45 percent will have received adequate prenatal care?

Solution: We can assume that the sampling distribution of $\hat{p}$ is approximately normally distributed with $\mu_{\hat{p}} = .51$ and $\sigma^2_{\hat{p}} = (.51)(.49)/200 = .00125$. We compute

$$z = \frac{.45 - .51}{\sqrt{.00125}} = \frac{-.06}{.0353} = -1.70$$

The area to the left of -1.70 under the standard normal curve is .0446. Therefore, $P(\hat{p} \leq .45) = P(z \leq -1.70) = .0446$. ■

EXERCISES

5.5.1 Smith et al. (A-5) performed a retrospective analysis of data on 782 eligible patients admitted with myocardial infarction to a 46-bed cardiac service facility. Of these patients, 248 (32 percent) reported a past myocardial infarction. Use .32 as the population proportion. Suppose 50 subjects are chosen at random from the population. What is the probability that over 40 percent would report previous myocardial infarctions?

5.5.2 In the study cited in Exercise 5.5.1, 13 percent of the patients in the study reported previous episodes of stroke or transient ischemic attack. Use 13 percent as the estimate of the prevalence of stroke or transient ischemic attack within the population. If 70 subjects are chosen at random from the population, what is the probability that 10 percent or less would report an incidence of stroke or transient ischemic attack?

5.5.3 In the same 1999–2000 NHANES (A-3) report cited in Example 5.5.1, researchers estimated that 64 percent of U.S. adults ages 20–74 were overweight or obese (overweight: BMI 25-29, obese: BMI 30 or greater). Use this estimate as the population proportion for U.S. adults ages 20–74. If 125 subjects are selected at random from the population, what is the probability that 70 percent or more would be found to be overweight or obese?

5.5.4 Gallagher et al. (A-6) reported on a study to identify factors that influence women's attendance at cardiac rehabilitation programs. They found that by 12 weeks post-discharge, only 64 percent of eligible women attended such programs. Using 64 percent as an estimate of the attendance percentage of all eligible women, find the probability that in a sample of 45 women selected at random from the population of eligible women less than 50 percent would attend programs.

5.5.5 Given a population in which $p = .6$ and a random sample from this population of size 100, find:

(a) $P(\hat{p} \geq .65)$ (b) $P(\hat{p} \leq .58)$

(c) $P(.56 \leq \hat{p} \leq .63)$