CHAPTER 7

Survey Sampling

7.1 Introduction

Resting on the probabilistic foundations of the preceding chapters, this chapter marks the beginning of our study of statistics by introducing the subject of survey sampling. As well as being of considerable intrinsic interest and practical utility, the development of the elementary theory of survey sampling serves to introduce several concepts and techniques that will recur and be amplified in later chapters.

Sample surveys are used to obtain information about a large population by examining only a small fraction of that population. Sampling techniques have been used in many fields, such as the following:

- Governments survey human populations; for example, the U.S. government conducts health surveys and census surveys.
- Sampling techniques have been extensively employed in agriculture to estimate such quantities as the total acreage of wheat in a state by surveying a sample of farms.
- The Interstate Commerce Commission has carried out sampling studies of rail and highway traffic. In one such study, records of shipments of household goods by motor carriers were sampled to evaluate the accuracy of preshipment estimates of charges, claims for damages, and other variables.
- In the practice of quality control, the output of a manufacturing process may be sampled in order to examine the items for defects.
- During audits of the financial records of large companies, sampling techniques may be used when examination of the entire set of records is impractical.

The sampling techniques discussed here are probabilistic in nature—each member of the population has a specified probability of being included in the sample, and the actual composition of the sample is random. Such techniques differ markedly from the type of sampling scheme in which particular population members are included in the sample because the investigator thinks they are typical in some way. Such a scheme may be effective in some situations, but there is no way mathematically to guarantee its unbiasedness (a term that will be precisely defined later) or to estimate the magnitude of any error committed, such as that arising from estimating the population mean by the sample mean. We will see that using a random sampling technique has a consequence that estimates can be guaranteed to be unbiased and probabilistic bounds on errors can be calculated. Among the advantages of using random sampling are the following:

- The selection of sample units at random is a guard against investigator biases, even unconscious ones.
- A small sample costs far less and is much faster to survey than a complete enumeration.
- The results from a small sample may actually be more accurate than those from a complete enumeration. The quality of the data in a small sample can be more easily monitored and controlled, and a complete enumeration may require a much larger, and therefore perhaps more poorly trained, staff.
- Random sampling techniques make possible the calculation of an estimate of the error due to sampling.
- In designing a sample, it is frequently possible to determine the sample size necessary to obtain a prescribed error level.

Peck et al. (2005) contains several interesting papers about applications of sampling.

7.2 Population Parameters

This section defines those numerical characteristics, or parameters, of the population that we will estimate from a sample. We will assume that the population is of size N and that associated with each member of the population is a numerical value of interest. These numerical values will be denoted by $x_1, x_2, \cdots, x_N$. The variable x_i may be a numerical variable such as age or weight, or it may take on the value 1 or 0 to denote the presence or absence of some characteristic such as gender. We will refer to the latter situation as the dichotomous case.

EXAMPLE **A** This is the first of many examples in this chapter in which we will illustrate ideas by using a study by Herkson (1976). The population consists of $N = 393$ short-stay hospitals. We will let x_i denote the number of patients discharged from the ith hospital during January 1968. A histogram of the population values is shown in Figure 7.1. The histogram was constructed in the following way: The number of hospitals that discharged 0–200, 201–400, . . . , 2801–3000 patients were graphed as horizontal lines above the respective intervals. For example, the figure indicates that about

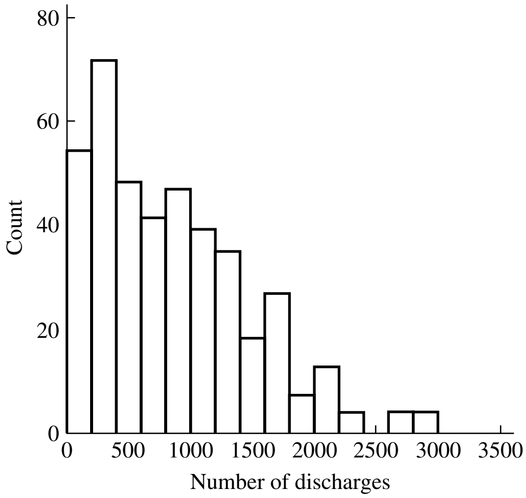

FIGURE **7.1** Histogram of the numbers of patients discharged during January 1968 from 393 short-stay hospitals.

40 hospitals discharged from 601 to 800 patients. The histogram is a convenient graphical representation of the distribution of the values in the population, being more quickly assimilated than would a list of 393 values. ■

We will be particularly interested in the **population mean,** or average,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

For the population of 393 hospitals, the mean number of discharges is 814.6. Note the location of this value in Figure 7.1. In the dichotomous case, where the presence or absence of a characteristic is to be determined, μ equals the proportion, p, of individuals in the population having the particular characteristic, because in the sum above, each x_i is either 0 or 1. The sum thus reduces to the number of 1s and when divided by N, gives the proportion, p.

The **population total** is

$$\tau = \sum_{i=1}^{N} x_i = N\mu$$

The total number of people discharged from the population of hospitals is $\tau = 320{,}138$. In the dichotomous case, the population total is the total number of members of the population possessing the characteristic of interest.

We will also need to consider the **population variance,**

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

A useful identity can be obtained by expanding the square in this equation:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - 2\mu\sum_{i=1}^{N} x_i + N\mu^2\right)\\ &= \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - 2N\mu^2 + N\mu^2\right)\\ &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\end{aligned}$$

In the dichotomous case, the population variance reduces to $p(1-p)$:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\\ &= p - p^2\\ &= p(1-p)\end{aligned}$$

Here we used the fact that because each x_i is 0 or 1, each x_i^2 is also 0 or 1.

The **population standard deviation** is the square root of the population variance and is used as a measure of how spread out, dispersed, or scattered the individual values are. The standard deviation is given in the same units (for example, inches) as are the population values, whereas the variance is given in those units squared. The variance of the discharges is 347,766, and the standard deviation is 589.7; examination of the histogram in Figure 7.1 makes it clear that the latter number is the more reasonable description of the spread of the population values.

7.3 Simple Random Sampling

The most elementary form of sampling is **simple random sampling** (s.r.s.): Each particular sample of size n has the same probability of occurrence; that is, each of the $\binom{N}{n}$ possible samples of size n taken without replacement has the same probability. We assume that sampling is done without replacement so that each member of the population will appear in the sample at most once. The actual composition of the sample is usually determined by using a table of random numbers or a random number generator on a computer. Conceptually, we can regard the population members as balls in an urn, a specified number of which are selected for inclusion in the sample at random and without replacement.

Because the composition of the sample is random, the sample mean is random. An analysis of the accuracy with which the sample mean approximates the population mean must therefore be probabilistic in nature. In this section, we will derive some statistical properties of the sample mean.

7.3.1 The Expectation and Variance of the Sample Mean

We will denote the sample size by n (n is less than N) and the values of the sample members by $X_1, X_2, \ldots, X_n$. It is important to realize that each X_i is a random variable. In particular, X_i is not the same as x_i: X_i is the value of the ith member of the sample, which is random and x_i is that of the ith member of the population, which is fixed.

We will consider the **sample mean,**

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

as an estimate of the population mean. As an estimate of the population total, we will consider

$$T = N\overline{X}$$

Properties of T will follow readily from those of $\overline{X}$. Since each X_i is a random variable, so is the sample mean; its probability distribution is called its **sampling distribution.** In general, any numerical value, or statistic, computed from a random sample is a random variable and has an associated sampling distribution. The sampling distribution of $\overline{X}$ determines how accurately $\overline{X}$ estimates μ; roughly speaking, the more tightly the sampling distribution is centered on μ, the better the estimate.

EXAMPLE **A** To illustrate the concept of a sampling distribution, let us look again at the population of 393 hospitals. In practice, of course, the population would not be known, and only one sample would be drawn. For pedagogical purposes here, we can consider the sampling distribution of the sample mean from this known population. Suppose, for example, that we want to find the sampling distribution of the mean of a sample of size 16. In principle, we could form all $\binom{393}{16}$ samples and compute the mean of each one—this would give the sampling distribution. But because the number of such samples is of the order 10^{33}, this is clearly not practical. We will thus employ a technique known as **simulation.** We can estimate the sampling distribution of the mean of a sample of size n by drawing many samples of size n, computing the mean of each sample, and then forming a histogram of the collection of sample means. Figure 7.2 shows the results of such a simulation for sample sizes of 8, 16, 32, and 64 with 500 replications for each sample size. Three features of Figure 7.2 are noteworthy:

1. All the histograms are centered about the population mean, 814.6.
2. As the sample size increases, the histograms become less spread out.
3. Although the shape of the histogram of population values (Figure 7.1) is not symmetric about the mean, the histograms in Figure 7.2 are more nearly so.

These features will be explained quantitatively. ■

As we have said, $\overline{X}$ is a random variable whose distribution is determined by that of the X_i. We thus examine the distribution of a single sample element, X_i. It should be noted that the following lemma holds whether sampling is with or without replacement.

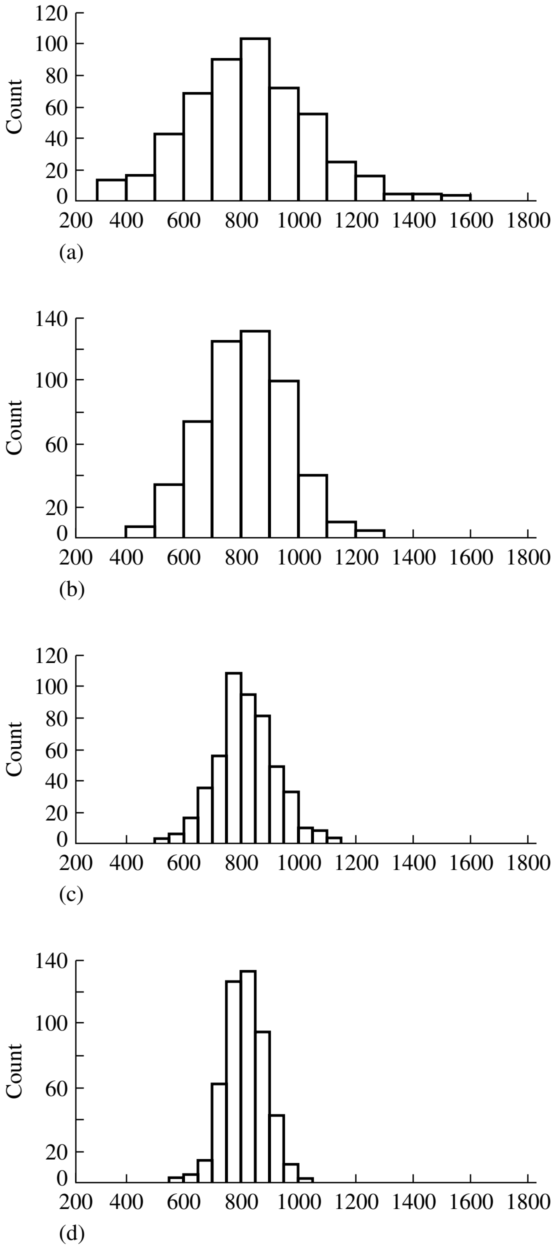

FIGURE **7.2** Histograms of the values of the mean number of discharges in 500 simple random samples from the population of 393 hospitals. Sample sizes: (a) $n = 8$, (b) $n = 16$, (c) $n = 32$, (d) $n = 64$.

We need to be careful about the values that the random variable X_i can assume. The i^{th} sample member is equally likely to be any of the N population members. If all the population values were distinct, we would then have $P(X_1 = x_j) = 1/N$. But the population values may not be distinct (for example, in the dichotomous case there are only two values, 0 and 1). If k members of the population have the same value ζ, then $P(X_i = \zeta) = k/N$. We use this construction in proving the following lemma.

LEMMA A

Denote the distinct values assumed by the population members by $\zeta_1, \zeta_2, \dots, \zeta_m$, and denote the number of population members that have the value ζ_j by n_j, $j = 1, 2, \dots, m$. Then X_i is a discrete random variable with probability mass function

$$P(X_i = \zeta_j) = \frac{n_j}{N}$$

Also,

$$E(X_i) = \mu$$
$$\mathrm{Var}(X_i) = \sigma^2$$

Proof

The only possible values that X_i can assume are $\zeta_1, \zeta_2, \dots, \zeta_m$. Since each member of the population is equally likely to be the ith member of the sample, the probability that X_i assumes the value ζ_j is thus n_j/N. The expected value of the random variable X_i is then

$$E(X_i) = \sum_{j=1}^{m} \zeta_j P(X_i = \zeta_j) = \frac{1}{N}\sum_{j=1}^{m} n_j \zeta_j = \mu$$

The last equation follows because n_j population members have the value ζ_j and the sum is thus equal to the sum of the values of all the population members. Finally,

$$\begin{aligned}\mathrm{Var}(X_i) &= E\left(X_i^2\right) - [E(X_i)]^2 \\ &= \frac{1}{N}\sum_{j=1}^{m} n_j \zeta_j^2 - \mu^2 \\ &= \sigma^2\end{aligned}$$

Here we have used the fact that $\sum_{i=1}^{N} x_i^2 = \sum_{j=1}^{m} n_j \zeta_j^2$ and the identity for the population variance derived in Section 7.2. ■

As a measure of the center of the sampling distribution, we will use $E(\overline{X})$. As a measure of the dispersion of the sampling distribution about this center, we will use the standard deviation of $\overline{X}$. The key results that will be obtained shortly are that the sampling distribution is centered at μ and that its spread is inversely proportional to the square root of the sample size, n. We first show that the sampling distribution is centered at μ.

THEOREM **A**

With simple random sampling, $E(\overline{X}) = \mu$.

Proof

Since, from Lemma A, $E(X_i) = \mu$, it follows from Theorem A in Section 4.1.2 that

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu \qquad \blacksquare$$

From Theorem A, we have the following corollary.

COROLLARY **A**

With simple random sampling, $E(T) = \tau$.

Proof

$$\begin{aligned} E(T) &= E(N\overline{X}) \\ &= NE(\overline{X}) \\ &= N\mu \\ &= \tau \end{aligned} \qquad \blacksquare$$

In the dichotomous case, $\mu = p$, and $\overline{X}$ is the proportion of the sample that possesses the characteristic of interest. In this case, $\overline{X}$ will be denoted by $\hat{p}$. We have shown that $E(\hat{p}) = p$.

It is important to keep in mind that $\overline{X}$ is random. The result $E(\overline{X}) = \mu$ can be interpreted to mean that "on the average" $\overline{X} = \mu$. In general, if we wish to estimate a population parameter, θ say, by a function $\hat{\theta}$ of the sample, $X_1, X_2, \ldots, X_n$, and $E(\hat{\theta}) = \theta$, whatever the value of θ may be, we say that $\hat{\theta}$ is **unbiased.** Thus, $\overline{X}$ and T are unbiased estimates of μ and τ. On average they are correct. We next investigate how variable they are, by deriving their variances and standard deviations. Section 4.2.1 introduced the concepts of bias and variance in the context of a model of measurement error, and these concepts are also relevant in this new context. In Chapter 4, it was shown that

$$\text{Mean squared error} = \text{variance} + \text{bias}^2$$

Since $\overline{X}$ and T are unbiased, their mean squared errors are equal to their variances.

We next find $\text{Var}(\overline{X})$. Since $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$, it follows from Corollary A of Section 4.3 that

$$\text{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}(X_i, X_j)$$

Suppose that sampling were done with replacement. Then the X_i would be independent, and for $i \neq j$ we would have $\text{Cov}(X_i, X_j) = 0$, whereas $\text{Cov}(X_i, X_i) = \text{Var}(X_i) = \sigma^2$. It would then follow that

$$\begin{aligned} \text{Var}\,\overline{X} &= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

and that the standard deviation of $\overline{X}$, also called its **standard error,** would be

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Sampling without replacement induces dependence among the X_i, which complicates this simple result. However, we will see that if the sample size n is small relative to the population size N, the dependence is weak and this simple result holds to a good approximation.

To find the variance of the sample mean in sampling without replacement we need to find $\text{Cov}(X_i, X_j)$ for $i \neq j$.

LEMMA **B**

For simple random sampling without replacement,

$$\text{Cov}(X_i, X_j) = -\sigma^2/(N-1) \qquad \text{if } i \neq j$$

Using the identity for covariance established at the beginning of Section 4.3,

$$\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$$

and

$$\begin{aligned} E(X_i X_j) &= \sum_{k=1}^{m}\sum_{l=1}^{m} \zeta_k \zeta_l P(X_i = \zeta_k \text{ and } X_j = \zeta_l) \\ &= \sum_{k=1}^{m} \zeta_k P(X_i = \zeta_k) \sum_{l=1}^{m} \zeta_l P(X_j = \zeta_l | X_i = \zeta_k) \end{aligned}$$

from the multiplication law for conditional probability. Now,

$$P(X_j = \zeta_l | X_i = \zeta_k) = \begin{cases} n_l/(N-1), & \text{if } k \neq l \\ (n_l - 1)/(N-1), & \text{if } k = l \end{cases}$$

Now if we express

$$\begin{aligned} \sum_{l=1}^{m} \zeta_l P(X_j = \zeta_l | X_i = \zeta_k) &= \sum_{l \neq k} \zeta_l \frac{n_l}{N-1} + \zeta_k \frac{n_k - 1}{N-1} \\ &= \sum_{l=1}^{m} \zeta_l \frac{n_l}{N-1} - \zeta_k \frac{1}{N-1} \end{aligned}$$

the expression for $E(X_i X_j)$ becomes

$$\sum_{k=1}^{m} \zeta_k \frac{n_k}{N} \left(\sum_{l=1}^{m} \zeta_l \frac{n_l}{N-1} - \frac{\zeta_k}{N-1} \right) = \frac{1}{N(N-1)} \left(\tau^2 - \sum_{k=1}^{m} \zeta_k^2 n_k \right)$$

$$= \frac{\tau^2}{N(N-1)} - \frac{1}{N(N-1)} \sum_{k=1}^{m} \zeta_k^2 n_k$$

$$= \frac{N\mu^2}{N-1} - \frac{1}{N-1}(\mu^2 + \sigma^2)$$

$$= \mu^2 - \frac{\sigma^2}{N-1}$$

Finally, subtracting $E(X_i)E(X_j) = \mu^2$ from the last equation, we have

$$\operatorname{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

for $i \neq j$. ■

(Alternative proofs of Lemma B are outlined in Problems 25 and 26 at the end of this chapter.) This lemma shows that X_i and X_j are not independent of each other for $i \neq j$, but that the covariance is very small for large values of N. We are now able to derive the following theorem.

THEOREM **B**

With simple random sampling,

$$\operatorname{Var}(\overline{X}) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$$

$$= \frac{\sigma^2}{n} \left(1 - \frac{n-1}{N-1} \right)$$

Proof

From Corollary A of Section 4.3,

$$\operatorname{Var}(\overline{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}(X_i) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \operatorname{Cov}(X_i, X_j)$$

$$= \frac{\sigma^2}{n} - \frac{1}{n^2} n(n-1) \frac{\sigma^2}{N-1}$$

After some algebra, this gives the desired result. ■

Notice that the variance of the sample mean in sampling without replacement differs from that in sampling with replacement by the factor

$$\left(1 - \frac{n-1}{N-1}\right)$$

which is called the **finite population correction.** The ratio n/N is called the **sampling fraction.** Frequently, the sampling fraction is very small, in which case the **standard error** (standard deviation) of $\overline{X}$ is

$$\sigma_{\overline{X}} \approx \frac{\sigma}{\sqrt{n}}$$

We see that, apart from the usually small finite population correction, the spread of the sampling distribution and therefore the precision of $\overline{X}$ are determined by the sample size (n) and not by the population size (N). As will be made more explicit later, the appropriate measure of the precision of the sample mean is its standard error, which is inversely proportional to the square root of the sample size. Thus, in order to double the accuracy, the sample size must be quadrupled. (You might examine Figure 7.2 with this in mind.) The other factor that determines the accuracy of the sample mean is the population standard deviation, σ. If σ is small, the population values are not very dispersed and a small sample will be fairly accurate. But if the values are widely dispersed, a much larger sample will be required in order to attain the same accuracy.

EXAMPLE **B** If the population of hospitals is sampled without replacement and the sample size is $n = 32$,

$$\begin{aligned}\sigma_{\overline{X}} &= \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}} \\ &= \frac{589.7}{\sqrt{32}}\sqrt{1 - \frac{31}{392}} \\ &= 104.2 \times .96 \\ &= 100.0\end{aligned}$$

Notice that because the sampling fraction is small, the finite population correction makes little difference. To see that $\sigma_{\overline{X}} = 100.0$ is a reasonable measure of accuracy, examine part (b) of Figure 7.2 and observe that the vast majority of sample means differed from the population mean (814) by less than two standard errors; i.e., the vast majority of sample means were in the interval (614, 1014). ■

EXAMPLE **C** Let us apply this result to the problem of estimating a proportion. In the population of hospitals, a proportion $p = .654$ had fewer than 1000 discharges. If this proportion were estimated from a sample as the sample proportion $\hat{p}$, the standard error of $\hat{p}$

could be found by applying Theorem B to this dichotomous case:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{1-\frac{n-1}{N-1}}$$

For example, for $n = 32$, the standard error of $\hat{p}$ is

$$\begin{aligned}\sigma_{\hat{p}} &= \sqrt{\frac{.654 \times .346}{32}}\sqrt{1-\frac{31}{392}}\\ &= .08\end{aligned}$$

■

The precision of the estimate of the population total does depend on the population size, N.

COROLLARY **B**

With simple random sampling,

$$\mathrm{Var}(T) = N^2\left(\frac{\sigma^2}{n}\right)\frac{N-n}{N-1}$$

Proof

Since $T = N\overline{X}$,

$$\mathrm{Var}(T) = N^2\ \mathrm{Var}(\overline{X})$$

■

7.3.2 Estimation of the Population Variance

A sample survey is used to estimate population parameters, and it is desirable also to assess and quantify the variability of the estimates. In the previous section, we saw how the standard error of an estimate may be determined from the sample size and the population variance. In practice, however, the population variance will not be known, but as we will show in this section, it can be estimated from the sample. Since the population variance is the average squared deviation from the population mean, estimating it by the average squared deviation from the sample mean seems natural:

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

The following theorem shows that this estimate is biased.

THEOREM **A**

With simple random sampling,

$$E(\hat{\sigma}^2) = \sigma^2 \left(\frac{n-1}{n}\right) \frac{N}{N-1}$$

Proof

Expanding the square and proceeding as in the identity for the population variance in Section 7.2, we find

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

Thus,

$$E(\hat{\sigma}^2) = \frac{1}{n} \sum_{i=1}^{n} E\left(X_i^2\right) - E(\overline{X}^2)$$

Now, we know that

$$\begin{aligned} E\left(X_i^2\right) &= \mathrm{Var}(X_i) + [E(X_i)]^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

Similarly, from Theorems A and B of Section 7.3.1,

$$\begin{aligned} E(\overline{X}^2) &= \mathrm{Var}(\overline{X}) + [E(\overline{X})]^2 \\ &= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) + \mu^2 \end{aligned}$$

Substituting these expressions for $E(X_i^2)$ and $E(\overline{X}^2)$ in the preceding equation for $E(\hat{\sigma}^2)$ gives the desired result. ■

Because $N > n$, it follows with a little algebra that

$$\frac{n-1}{n} \frac{N}{N-1} < 1$$

so that $E(\hat{\sigma}^2) < \sigma^2$; $\hat{\sigma}^2$ thus tends to underestimate σ^2. From Theorem A, we see that an unbiased estimate of σ^2 may be obtained by multiplying $\hat{\sigma}^2$ by the factor $n(N-1)/[(n-1)N]$. Thus, an unbiased estimate of σ^2 is $\frac{1}{n-1}(1-\frac{1}{N}) \sum_{i=1}^{n} (X_i - \overline{X})^2$. We also have the following corollary.

COROLLARY **A**

An unbiased estimate of $\mathrm{Var}(\overline{X})$ is

$$\begin{aligned} s_{\overline{X}}^2 &= \frac{\hat{\sigma}^2}{n}\left(\frac{n}{n-1}\right)\left(\frac{N-1}{N}\right)\left(\frac{N-n}{N-1}\right) \\ &= \frac{s^2}{n}\left(1-\frac{n}{N}\right) \end{aligned}$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Proof

Since

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

an unbiased estimate of $\mathrm{Var}(\overline{X})$ may be obtained by substituting in an unbiased estimate of σ^2. Algebra then yields the desired result. ■

Similarly, an unbiased estimate of the variance of T, the estimator of the population total, is

$$s_T^2 = N^2 s_{\overline{X}}^2$$

For the dichotomous case, in which each X_i is 0 or 1, note that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \\ &= \hat{p}(1-\hat{p}) \end{aligned}$$

Therefore,

$$s^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

Thus, as a special case of Corollary A, we have the following corollary.

COROLLARY **B**

An unbiased estimate of $\mathrm{Var}(\hat{p})$ is

$$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)$$ ■

In many cases, the sampling fraction, n/N, is small and may be neglected. Furthermore, it often makes little difference whether $n-1$ or n is used as the divisor.

The quantities $s_{\overline{X}}$, s_T, and $s_{\hat{p}}$ are called **estimated standard errors.** If we knew them, the actual standard errors, $\sigma_{\overline{X}}$, σ_T and $\sigma_{\hat{p}}$, would be used to gauge the accuracy of the estimates $\overline{X}$, T and $\hat{p}$. If they are not known, which is the typical case, the estimated standard errors are used in their place.

EXAMPLE **A** A simple random sample of 50 of the 393 hospitals was taken. From this sample, $\overline{X} = 938.5$ (recall that, in fact, $\mu = 814.6$) and $s = 614.53$ ($\sigma = 590$). An estimate of the variance of $\overline{X}$ is

$$s_{\overline{X}}^2 = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = 6592$$

The estimated standard error of $\overline{X}$ is

$$s_{\overline{X}} = 81.19$$

(Note that the true value is $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{50}}\sqrt{1 - \frac{49}{392}} = 78$.) This estimated standard error gives a rough idea of how accurate the value of $\overline{X}$ is; in this case, we see that the magnitude of the error is of the order 80, as opposed to 8 or 800, say. In fact, the error was 123.9, or about 1.5 $s_{\overline{X}}$. ■

EXAMPLE **B** From the same sample, the estimate of the total number of discharges in the population of hospitals is

$$T = N\overline{X} = 368{,}831$$

Recall that the true value of the population total is 320,139. The estimated standard error of T is

$$s_T = Ns_{\overline{X}} = 31{,}908$$

Again, this estimated standard error can be used as a rough gauge of the estimation error. ■

EXAMPLE **C** Let p be the proportion of hospitals that had fewer than 1000 discharges—that is, $p = .654$. In the sample of Example A, 26 of 50 hospitals had fewer than 1000 discharges, so

$$\hat{p} = \frac{26}{50} = .52$$

The variance of $\hat{p}$ is estimated by

$$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right) = .0045$$

Thus, the estimated standard error of $\hat{p}$ is

$$s_{\hat{p}} = .067$$

Crudely, this tells us that the error of $\hat{p}$ is in the second or first decimal place—that we are probably not so fortunate as to have an error only in the third decimal place. In fact, the error was .134 or about $2 \times s_{\hat{p}}$. ■

These examples show how, in simple random sampling, we can not only form estimates of unknown population parameters, but can also gauge the likely size of the errors of the estimates, by estimating their standard errors from the data in the sample.

We have covered a lot of ground, and the presence of the finite population correction complicates the expressions we have derived. It is thus useful to summarize our results in the following table:

Population Parameter	Estimate	Variance of Estimate	Estimated Variance
μ	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$	$s_{\overline{X}}^2 = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$
p	$\hat{p} =$ sample proportion	$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$	$s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right)$
τ	$T = N\overline{X}$	$\sigma_T^2 = N^2\sigma_{\overline{X}}^2$	$s_T^2 = N^2 s_{\overline{X}}^2$
σ^2	$\left(1 - \frac{1}{N}\right)s^2$		

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

The square roots of the entries in the third column are called *standard errors,* and the square roots of the entries in the fourth column are called *estimated standard errors.* The former depend on unknown population parameters, so the latter are used to gauge the accuracy of the parameter estimates. When the population is large relative to the sample size, the finite population correction can be ignored, simplifying the preceding expressions.

7.3.3 The Normal Approximation to the Sampling Distribution of $\overline{X}$

We have found the mean and the standard deviation of the sampling distribution of $\overline{X}$. Ideally, we would like to know the sampling distribution, since it would tell us everything we could hope to know about the accuracy of the estimate. Without knowledge of the population itself, however, we cannot determine the sampling distribution. In this section, we will use the central limit theorem to deduce an approximation to the sampling distribution—the normal, or Gaussian, distribution. This approximation will be used to find probabilistic bounds for the estimation error.

In Section 5.3, we considered a sequence of independent and identically distributed (i.i.d.) random variables, $X_1, X_2, \ldots$ having the common mean and variance μ and σ^2. The sample mean of $X_1, X_2, \ldots, X_n$ is

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This sample mean has the properties

$$E(\overline{X}_n) = \mu$$

and

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

The central limit theorem says that, for a fixed number z,

$$P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z) \qquad \text{as } n \to \infty$$

where Φ is the cumulative distribution function of the standard normal distribution. Using a more compact and suggestive notation, we have

$$P\left(\frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}} \leq z\right) \to \Phi(z)$$

The context of survey sampling is not exactly like that of the central limit theorem as stated above—as we have seen, in sampling without replacement, the X_i are not independent of each other, and it makes no sense to have n tend to infinity while N remains fixed. But other central limit theorems have been proved that are appropriate to the sampling context. These show that if n is large, but still small relative to N, then $\overline{X}_n$, the mean of a simple random sample, is approximately normally distributed.

To demonstrate the use of the central limit theorem, we will apply it to approximate $P(|\overline{X} - \mu| \leq \delta)$, the probability that the error made in estimating μ by $\overline{X}$ is less than some constant δ

$$\begin{aligned} P(|\overline{X} - \mu| \leq \delta) &= P(-\delta \leq \overline{X} - \mu \leq \delta) \\ &= P\left(-\frac{\delta}{\sigma_{\overline{X}}} \leq \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \leq \frac{\delta}{\sigma_{\overline{X}}}\right) \\ &\approx \Phi\left(\frac{\delta}{\sigma_{\overline{X}}}\right) - \Phi\left(-\frac{\delta}{\sigma_{\overline{X}}}\right) \\ &= 2\Phi\left(\frac{\delta}{\sigma_{\overline{X}}}\right) - 1 \end{aligned}$$

since $\Phi(-z) = 1 - \Phi(z)$, from the symmetry of the standard normal distribution about zero.

EXAMPLE A Let us again consider the population of 393 hospitals. The standard deviation of the mean of a sample of size $n = 64$ is, using the finite population correction,

$$\begin{aligned} \sigma_{\overline{X}} &= \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}} \\ &= \frac{589.7}{8}\sqrt{1 - \frac{63}{392}} = 67.5 \end{aligned}$$

We can use the central limit theorem to approximate the probability that the sample mean differs from the population mean by more than 100 in absolute value; i.e.,

$P(|\overline{X} - \mu| > 100)$. First, from the symmetry of the normal distribution,

$$P(|\overline{X} - \mu| > 100) \approx 2P(\overline{X} - \mu > 100)$$

and

$$\begin{aligned} P(\overline{X} - \mu > 100) &= 1 - P(\overline{X} - \mu < 100) \\ &= 1 - P\left(\frac{\overline{X} - \mu}{\sigma_{\overline{X}}} < \frac{100}{\sigma_{\overline{X}}}\right) \\ &\approx 1 - \Phi\left(\frac{100}{67.5}\right) \\ &= .069 \end{aligned}$$

Thus the probability that the sample mean differs from the population mean by more than 100 is approximately .14. In fact, among the 500 samples of size 64 in Example A in Section 7.3.1, 82, or 16.4%, differed by more than 100 from the population mean. Similarly, the central limit theorem approximation gives .026 as the probability of deviations of more than 150 from the population mean. In the simulation in Example A in Section 7.3.1, 11 of 500, or 2.2%, differed by more than 150. If we are not too finicky, the central limit theorem gives us reasonable and useful approximations. ■

EXAMPLE **B** For a sample of size 50, the standard error of the sample mean number of discharges is

$$\sigma_{\overline{X}} = 78$$

For the particular sample of size 50 discussed in Example A in Section 7.3.2, we found $\overline{X} = 938.35$, so $\overline{X} - \mu = 123.9$. We now calculate an approximation of the probability of an error this large or larger:

$$\begin{aligned} P(|\overline{X} - \mu| \geq 123.9) &= 1 - P(|\overline{X} - \mu| < 123.9) \\ &\approx 1 - \left[2\Phi\left(\frac{123.9}{78}\right) - 1\right] \\ &= 2 - 2\Phi(1.59) \\ &= .11 \end{aligned}$$

Thus, we can expect an error this large or larger to occur about 11% of the time. ■

EXAMPLE **C** In Example C in Section 7.3.2, we found from the sample of size 50 an estimate $\hat{p} = .52$ of the proportion of hospitals that discharged fewer than 1000 patients; in fact, the actual proportion in the population is .65. Thus, $|\hat{p} - p| = .13$. What is the probability that an estimate will be off by an amount this large or larger?

We have

$$\begin{aligned} \sigma_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}}\sqrt{1 - \frac{n-1}{N-1}} \\ &= .068 \times .94 = .064 \end{aligned}$$

We can therefore calculate

$$\begin{aligned} P(|p - \hat{p}| > .13) &= 1 - P(|p - \hat{p}| \le .13) \\ &= 1 - P\left(\frac{|p - \hat{p}|}{\sigma_{\hat{p}}} \le \frac{.13}{\sigma_{\hat{p}}}\right) \\ &\approx 2[1 - \Phi(2.03)] = .04 \end{aligned}$$

We see that the sample was rather "unlucky"—an error this large or larger would occur only about 4% of the time. ■

We can now derive a **confidence interval** for the population mean, μ. A confidence interval for a population parameter, θ, is a random interval, calculated from the sample, that contains θ with some specified probability. For example, a 95% confidence interval for μ is a random interval that contains μ with probability .95; if we were to take many random samples and form a confidence interval from each one, about 95% of these intervals would contain μ. If the coverage probability is $1 - \alpha$, the interval is called a $100(1 - \alpha)$% confidence interval. Confidence intervals are frequently used in conjunction with point estimates to convey information about the uncertainty of the estimates.

For $0 \le \alpha \le 1$, let $z(\alpha)$ be that number such that the area under the standard normal density function to the right of $z(\alpha)$ is α (Figure 7.3). Note that the symmetry of the standard normal density function about zero implies that $z(1 - \alpha) = -z(\alpha)$. If Z follows a standard normal distribution, then, by definition of $z(\alpha)$,

$$P(-z(\alpha/2) \le Z \le z(\alpha/2)) = 1 - \alpha$$

From the central limit theorem, $(\overline{X} - \mu)/\sigma_{\overline{X}}$ has approximately a standard normal distribution, so

$$P\left(-z(\alpha/2) \le \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \le z(\alpha/2)\right) \approx 1 - \alpha$$

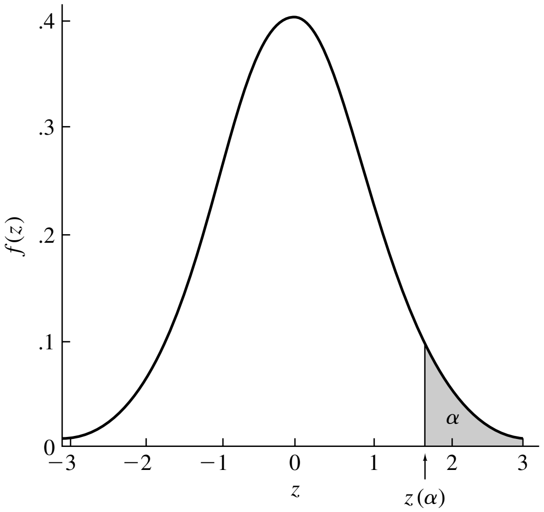

FIGURE 7.3 A standard normal density showing α and $z(\alpha)$.

Elementary manipulation of the inequalities gives

$$P(\overline{X} - z(\alpha/2)\sigma_{\overline{X}} \leq \mu \leq \overline{X} + z(\alpha/2)\sigma_{\overline{X}}) \approx 1 - \alpha$$

That is, the probability that μ lies in the interval $\overline{X} \pm z(\alpha/2)\sigma_{\overline{X}}$ is approximately $1 - \alpha$. The interval is thus called a $100(1 - \alpha)\%$ **confidence interval.** It is important to understand that this interval is random and that the preceding equation states that the probability that this random interval covers μ is $1 - \alpha$. In practice, α is assigned a small value, such as .1, .05, or .01, so that the probability that the interval covers μ will be large. Also, since the population variance is typically not known, $s_{\overline{X}}$ is substituted for $\sigma_{\overline{X}}$. For large samples, it can be shown that the effect of this substitution is practically negligible. It is impossible to give a precise answer to the question "How large is large?" As a rule of thumb, a value of n greater than 25 or 30 is usually adequate.

To illustrate the concept of a confidence interval, 20 samples each of size $n = 25$ were drawn from the population of hospital discharges. From each of these 20 samples, an approximate 95% confidence interval for μ, the mean number of discharges, was computed. These 20 confidence intervals are displayed as vertical lines in Figure 7.4; the dashed line in the figure is drawn at the true value, $\mu = 814.6$. Notice that it so

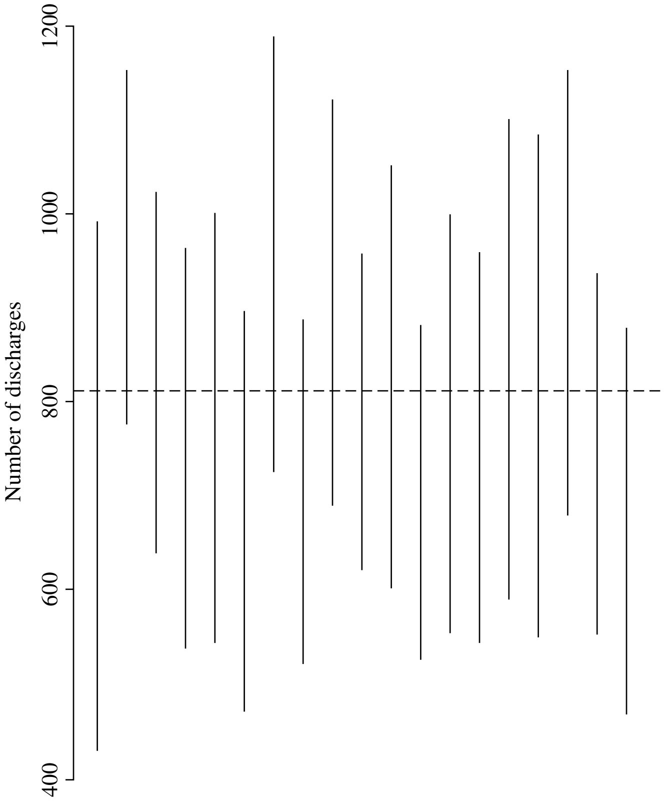

FIGURE **7.4** Vertical lines are 20 approximate 95% confidence intervals for μ. The horizontal line is the true value of μ.

happened that all the confidence intervals included μ; since these are 95% intervals, on the average 5%, or 1 out of 20, would not include μ.

The following example illustrates the procedure for calculating confidence intervals.

EXAMPLE **D** A particular area contains 8000 condominium units. In a survey of the occupants, a simple random sample of size 100 yields the information that the average number of motor vehicles per unit is 1.6, with a sample standard deviation of .8. The estimated standard error of $\overline{X}$ is thus

$$\begin{aligned} s_{\overline{X}} &= \frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \\ &= \frac{.8}{10}\sqrt{1-\frac{100}{8000}} \\ &= .08 \end{aligned}$$

Note that the finite population correction makes almost no difference. Since $z(.025) = 1.96$, a 95% confidence interval for the population average is $\overline{X} \pm 1.96 s_{\overline{X}}$, or (1.44, 1.76).

An estimate of the total number of motor vehicles is $T = 8000 \times 1.6 = 12{,}800$. The estimated standard error of T is

$$s_T = N s_{\overline{X}} = 640$$

A 95% confidence interval for the total number of motor vehicles is $T \pm 1.96 s_T$, or (11,546, 14,054).

In the same survey, 12% of the respondents said they planned to sell their condos within the next year; $\hat{p} = .12$ is an estimate of the population proportion p. The estimated standard error is

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\sqrt{1-\frac{100}{8000}} = .03$$

A 95% confidence interval for p is $\hat{p} \pm 1.96 s_{\hat{p}}$, or (.06, .18).

The total number of owners planning to sell is estimated as $T = N\hat{p} = 960$. The estimated standard error of T is $s_T = N s_{\hat{p}} = 240$. A 95% confidence interval for the number in the population planning to sell is $T \pm 1.96 s_T$, or (490, 1430). The proper interpretation of this interval, (490, 1430), is a little subtle. We cannot state that the probability is 0.95 and that the number of owners planning to sell is between 490 and 1430, because that number is either in this interval or not. What is true is that 95% of intervals formed in this way will contain the true number in the long run. This interval is like one of those shown in Figure 7.4; in the long run, 95% of those intervals will contain the true number of discharges, but in the figure any particular interval either does or doesn't contain the true number. ■

The width of a confidence interval is determined by the sample size n and the population standard deviation σ. If σ is known approximately, perhaps from earlier samples of the population, n can be chosen so as to obtain a confidence interval close to some desired length. Such analysis is usually an important aspect of planning the design of a sample survey.

EXAMPLE **E** The interval for the total number of owners planning to sell in Example D might be considered too wide for practical purposes; reducing its width would require a larger sample size. Suppose that an interval with a half-width of 200 is desired. Neglecting the finite population correction, the half-width is

$$1.96 s_T = 1.96 N \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = \frac{5095}{\sqrt{n-1}}$$

Setting the last expression equal to 200 and solving for n yields $n = 650$ as the necessary sample size. ■

Let us summarize: The fundamental result of this section is that the sampling distribution of the sample mean is approximately Gaussian. This approximation can be used to quantify the error committed in estimating the population mean by the sample mean, thus giving us a good understanding of the accuracy of estimates produced by a simple random sample. We next introduced the idea of a confidence interval, a random interval that contains a population parameter with a specified probability and thus provides an assessment of the accuracy of the corresponding estimate of that parameter. We have seen in our examples that the width of the confidence interval is a multiple of the estimated standard deviation of the estimate; for example, a confidence interval for μ is $\overline{X} \pm k s_{\overline{X}}$, where the constant k depends on the coverage probability of the interval.

7.4 Estimation of a Ratio

The foundations of the theory of survey sampling have been laid in the preceding sections on simple random sampling. This and the next section build on that foundation, developing some advanced topics in survey sampling.

In this section, we consider the estimation of a ratio. Suppose that for each member of a population, two values, x and y, may be measured. The ratio of interest is

$$r = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x}$$

Ratios arise frequently in sample surveys; for example, if households are sampled, the following ratios might be calculated:

- If y is the number of unemployed males aged 20–30 in a household and x is the number of males aged 20–30 in a household, then r is the proportion of unemployed males aged 20–30.

- If y is weekly food expenditure and x is number of inhabitants, then r is weekly food cost per inhabitant.
- If y is the number of motor vehicles and x is the number of inhabitants of driving age, then r is the number of motor vehicles per inhabitant of driving age.

In a survey of farms, y might be the acres of wheat planted and x the total acreage. In an inventory audit, y might be the audited value of an item and x the book value.

In this section, we first consider directly the problem of estimating a ratio. Later, we will use the estimation of a ratio as a technique for estimating μ_y. We will produce a new estimate, the ratio estimate, which we will compare to the ordinary estimate, $\overline{Y}$.

Before continuing, we note the elementary but sometimes overlooked fact that

$$r \neq \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{x_i}$$

Suppose that a sample is drawn consisting of the pairs (X_i, Y_i); the natural estimate of r is $R = \overline{Y}/\overline{X}$. We wish to derive expressions for $E(R)$ and $\text{Var}(R)$, but since R is a nonlinear function of the random variables $\overline{X}$ and $\overline{Y}$, we cannot do this in closed form. We will therefore employ the approximate methods of Section 4.6.

In order to calculate the approximate variance of R, we need to know $\text{Var}(\overline{X})$, $\text{Var}(\overline{Y})$, and $\text{Cov}(\overline{X}, \overline{Y})$. The first two quantities we know from Theorem B of Section 7.3.1. For the last quantity, we define the **population covariance** of x and y to be

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)$$

It can then be shown, in a manner entirely analogous to the proof of Theorem B in Section 7.3.1, that

$$\text{Cov}(\overline{X}, \overline{Y}) = \frac{\sigma_{xy}}{n}\left(1 - \frac{n-1}{N-1}\right)$$

From Example C in Section 4.6, we have the following theorem.

THEOREM A

With simple random sampling, the approximate variance of $R = \overline{Y}/\overline{X}$ is

$$\begin{aligned}\text{Var}(R) &\approx \frac{1}{\mu_x^2}\left(r^2\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2 - 2r\sigma_{\overline{X}\overline{Y}}\right) \\ &= \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}\right) \qquad \blacksquare\end{aligned}$$

The **population correlation coefficient** is defined as

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

and is used as a measure of the strength of the linear relationship between the x and y values in the population. It can be shown that $-1 \leq \rho \leq 1$; large values of ρ

indicate a strong positive relationship between x and y, and small values indicate a strong negative relationship. (See Figure 4.7 for some illustrations of correlation.) The equation in Theorem A can be expressed in terms of the population correlation coefficient as follows:

$$\mathrm{Var}(R) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y\right)$$

From this expression, we see that strong correlation of the same sign as r decreases the variance. We also note that the variance is affected by the size of μ_x—if μ_x is small, the variance is large, essentially because small values of $\overline{X}$ in the ratio $R = \overline{Y}/\overline{X}$ cause R to fluctuate wildly.

We now consider the approximate expectation of R. From Example C in Section 4.6 and the preceding calculations, we have the following theorem.

THEOREM **B**

With simple random sampling, the expectation of R is given approximately by

$$E(R) \approx r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r\sigma_x^2 - \rho\sigma_x\sigma_y\right) \qquad \blacksquare$$

From the equation in Theorem B, we see that strong correlation of the same sign as r decreases the bias and that the bias is large if μ_x is small. Furthermore, note that the bias is of the order $1/n$, so its contribution to the mean squared error is of the order $1/n^2$. In comparison, the contribution of the variance is of the order $1/n$. Therefore, for large samples, the bias is negligible compared to the standard error of the estimate.

For large samples, truncating the Taylor series after the linear term provides a good approximation, since the deviations $\overline{X} - \mu_X$ and $\overline{Y} - \mu_Y$ are likely to be small. To this order of approximation, R is expressed as a linear combination of $\overline{X}$ and $\overline{Y}$, and an argument based on the central limit theorem can be used to show that R is approximately normally distributed. Approximate confidence intervals can thus be formed for r by using the normal distribution.

In order to estimate the standard error of R, we substitute R for r in the formula of Theorem A. The x and y population variances are estimated by s_x^2 and s_y^2. The population covariance is estimated by

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_iY_i - n\overline{X}\overline{Y}\right) \end{aligned}$$

(as can be seen by expanding the product), and the population correlation is estimated by

$$\hat{\rho} = \frac{s_{xy}}{s_x s_y}$$

The estimated variance of R is thus

$$s_R^2 = \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\overline{X}^2}(R^2 s_x^2 + s_y^2 - 2Rs_{xy})$$

An approximate $100(1-\alpha)\%$ confidence interval for r is $R \pm z(\alpha/2)s_R$.

EXAMPLE **A** Suppose that 100 people who recently bought houses are surveyed, and the monthly mortgage payment and gross income of each buyer are determined. Let y denote the mortgage payment and x the gross income. Suppose that

$$\begin{aligned} \overline{X} &= \$3100 & \overline{Y} &= \$868 \\ s_y &= \$250 & s_x &= \$1200 \\ \hat{\rho} &= .85 & R &= .28 \end{aligned}$$

Neglecting the finite population correction, the estimated standard error of R is

$$\begin{aligned} s_R &= \frac{1}{10}\left(\frac{1}{3100}\right)\sqrt{.28^2 \times 1200^2 + 250^2 - 2 \times .28 \times .85 \times 250 \times 1200} \\ &= .006 \end{aligned}$$

An approximate 95% confidence interval for r is $.28 \pm (1.96) \times (.006)$, or $.28 \pm .012$. Note that the high correlation between x and y causes the standard error of R to be small. We can use the observed values for the variances, covariances, and means to gauge the order of magnitude of the bias by substituting them in place of the population parameters in the formula of Theorem B. Doing so, and again neglecting the finite population correction, gives the value .00015 for the bias, which is negligible relative to s_R. Note that the large value of $\overline{X}$ and the large positive correlation coefficient cause the bias to be small. ■

Ratios may also be used as tools for estimating population means and totals. To illustrate the concept, we return to the example of hospital discharges. For this population, the number of beds in each hospital is also known; let us denote the number of beds in the ith hospital by x_i and the number of discharges by y_i. Suppose that all the x_i are known, perhaps from an earlier enumeration, before a sample has been taken to estimate the number of discharges, and that we would like to take advantage of this information. One way to do this is to form a **ratio estimate** of μ_y:

$$\overline{Y}_R = \frac{\mu_x}{\overline{X}}\overline{Y} = \mu_x R$$

where $\overline{X}$ is the average number of beds and $\overline{Y}$ is the average number of discharges in the sample. The idea is fairly simple: We expect x_i and y_i to be closely related in the population, since a hospital with a large number of beds should tend to have a large number of discharges. This is borne out by Figure 7.5, a scatterplot of the number of discharges versus the number of beds. If $\overline{X} < \mu_x$, the sample underestimates the number of beds and probably the number of discharges as well; multiplying $\overline{Y}$ by $\mu_x/\overline{X}$ increases $\overline{Y}$ to $\overline{Y}_R$.

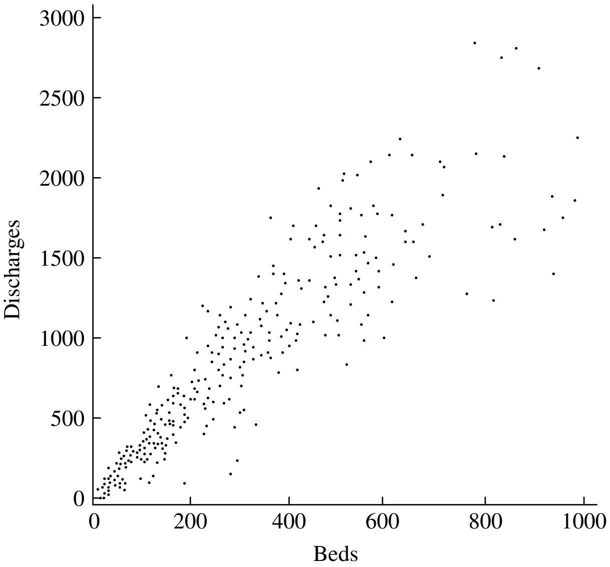

FIGURE **7.5** Scatterplot of the number of discharges versus the number of beds for the 393 hospitals.

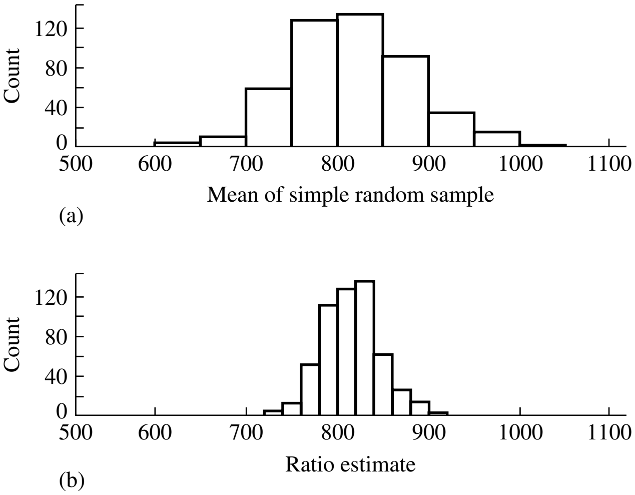

FIGURE **7.6** (a) A histogram of the means of 500 simple random samples of size 64 from the population of discharges; (b) a histogram of the values of 500 ratio estimates of the mean number of discharges from samples of size 64.

To see how this ratio estimate works in practice, it was simulated from 500 samples of size 64 from the population of hospitals. The histogram of the results is shown in Figure 7.6 along with the histogram of the means of 500 simple random samples of size 64. The comparison shows dramatically how effective the ratio estimate is at reducing variability.

Two more examples will illustrate the scope of the ratio estimation method.

EXAMPLE **B** Suppose that we want to estimate the total number of unemployed males aged 20–30 from a sample of households and that we know τ_x, the total number of males aged 20–30, from census data. The ratio estimate is

$$T_R = \tau_x \frac{\overline{Y}}{\overline{X}}$$

where $\overline{Y}$ is the average number of unemployed males aged 20–30 per household in the sample, and $\overline{X}$ is the sample average number of males aged 20–30 per household. ■

EXAMPLE **C** A sample of items in an inventory is taken to estimate the total value of the inventory. Let Y_i be the audited value of the ith sample item, and let X_i be its book value. We assume that τ_x, the total book value of the inventory, is known, and we estimate the total audited value by

$$T_R = \tau_x \frac{\overline{Y}}{\overline{X}}$$ ■

We will now analyze the observed success of the ratio estimate. Since $\overline{Y}_R = \mu_X R$, $\mathrm{Var}(\overline{Y}_R) = \mu_X^2 \mathrm{Var}(R)$. From Theorem A, we thus have the following.

COROLLARY **A**

The approximate variance of the ratio estimate of μ_y is

$$\mathrm{Var}(\overline{Y}_R) \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\rho\sigma_x\sigma_y\right)$$ ■

Similarly, from Theorem B, we have another corollary.

COROLLARY **B**

The approximate bias of the ratio estimate of μ_y is

$$E(\overline{Y}_R) - \mu_Y \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x}\left(r\sigma_x^2 - \rho\sigma_x\sigma_y\right)$$ ■

When will the ratio estimate Y_R be better than the ordinary estimate $\overline{Y}$? In the following, the finite population correction is neglected for simplicity. Since the variance of the ordinary estimate $\overline{Y}$ is

$$\mathrm{Var}(\overline{Y}) = \frac{\sigma_y^2}{n}$$

the ratio estimate has a smaller variance if

$$r^2\sigma_x^2 - 2r\rho\sigma_x\sigma_y < 0$$

or (provided $r > 0$, for example)

$$2\rho\sigma_y > r\sigma_x$$

Letting $C_x = \sigma_x/\mu_x$ and $C_y = \sigma_y/\mu_y$, this last inequality is equivalent to

$$\rho > \frac{1}{2}\left(\frac{C_x}{C_y}\right)$$

C_x and C_y are called **coefficients of variation** and give the standard deviation as a proportion of the mean. (Coefficients of variation are often more meaningful than standard deviations. For example, a standard deviation of 10 means one thing if the true value of the quantity being measured is 100 and something entirely different if the true value is 10,000.)

In order to assess the accuracy of $\overline{Y}_R$, Var($\overline{Y}_R$) can be estimated from the sample.

COROLLARY **C**

The variance of $\overline{Y}_R$ can be estimated by

$$s_{\overline{Y}_R}^2 = \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\left(R^2 s_x^2 + s_y^2 - 2Rs_{xy}\right)$$

and an approximate $100(1-\alpha)\%$ confidence interval for μ_y is $(\overline{Y}_R \pm z(\frac{\alpha}{2})s_{\overline{Y}_R})$. ■

EXAMPLE **D** For the population of 393 hospitals, we have

$$\begin{aligned} \mu_x &= 274.8 & \sigma_x &= 213.2 \\ \mu_y &= 814.6 & \sigma_y &= 589.7 \\ r &= 2.96 & \rho &= .91 \end{aligned}$$

Thus,

$$\begin{aligned} \text{Var}(\overline{Y}_R) &\approx \frac{1}{n}(2.96^2 \times 213.2^2 + 589.7^2 - 2 \times 2.96 \times .91 \times 213.2 \times 589.7) \\ &= \frac{68{,}697.4}{n} \end{aligned}$$

and

$$\sigma_{\overline{Y}_R} \approx \frac{262.1}{\sqrt{n}}$$

Including the finite population correction, the linearized approximation predicts that, with $n = 64$,

$$\sigma_{\overline{Y}_R} = \frac{1}{8}(262.1)\sqrt{1 - \frac{63}{392}} = 30.0$$

The actual standard deviation of the 500 sample values displayed in Figure 7.6 is 29.9, which is remarkably close. The mean of the 500 values is 816.2, compared to the population mean of 814.6; the slight apparent bias is consistent with Corollary B.

In contrast, the standard deviation of $\overline{Y}$ from a simple random sample of size $n = 64$ is

$$\begin{aligned}\sigma_{\overline{Y}} &= \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}}\\ &= \frac{589.7}{8}\sqrt{1 - \frac{63}{329}}\\ &= 66.3\end{aligned}$$

The comparison of $\sigma_{\overline{Y}}$ to $\sigma_{\overline{Y}_R}$ is consistent with the substantial reduction in variability accomplished by using a ratio estimate of μ_y shown in Figure 7.6.

The following is another way of interpreting this comparison. If a simple random sample of size n_1 is taken, the variance of the estimate is $\mathrm{Var}(\overline{Y}) = 589.7^2/n_1$. A ratio estimate from a sample of size n_2 will have the same variance if

$$\frac{262.1^2}{n_2} = \frac{589.7^2}{n_1}$$

or

$$n_2 = n_1\left(\frac{262.1}{589.7}\right)^2 = .1975n_1$$

Thus, in this case, we can obtain the same precision from a ratio estimate *using a sample about 80% smaller* than the simple random sample. Note that this comparison neglects the bias of the ratio estimate, which is justifiable in this case because the bias is quite small. Here is a case in which a biased estimate performs substantially better than an unbiased estimate, the bias being quite small and the reduction in variance being quite large. ■

7.5 Stratified Random Sampling

7.5.1 Introduction and Notation

In stratified random sampling, the population is partitioned into subpopulations, or **strata,** which are then independently sampled. The results from the strata are then combined to estimate population parameters, such as the mean.

Following are some examples that suggest the range of situations in which stratification is natural:

- In auditing financial transactions, the transactions may be grouped into strata on the basis of their nominal values. For example, high-value, medium-value, and low-value strata might be formed.
- In samples of human populations, geographical areas often form natural strata.
- In a study of records of shipments of household goods by motor carriers, the carriers were grouped into three strata: large carriers, medium carriers, and small carriers.

Stratified samples are used for a variety of reasons. We are often interested in obtaining information about each of a number of natural subpopulations in addition to information about the population as a whole. The subpopulations might be defined by geographical areas or age groups. In an industrial application in which the population consists of items produced by a manufacturing process, relevant subpopulations might consist of items produced during different shifts or from different lots of raw material. The use of a stratified random sample guarantees a prescribed number of observations from each subpopulation, whereas the use of a simple random sample can result in underrepresentation of some subpopulations. A second reason for using stratification is that, as will be shown below, the stratified sample mean can be considerably more precise than the mean of a simple random sample, especially if the population members within each stratum are relatively homogeneous and if there is considerable variation between strata.

In the next section, properties of the stratified sample mean are derived. Since a simple random sample is taken within each stratum, the results will follow easily from the derivations of earlier sections. The section after that takes up the problem of how to allocate the total number of observations, n, among the various strata. Comparisons will be made of the efficiencies of different allocation schemes and also of the precisions of these allocation schemes relative to that of a simple random sample of the same total size.

7.5.2 Properties of Stratified Estimates

Suppose there are L strata in all. Let the number of population elements in stratum 1 be denoted by N_1, the number in stratum 2 be N_2, etc. The total population size is $N = N_1 + N_2 + \ldots + N_L$. The population mean and variance of the lth stratum are denoted by μ_l and σ_l^2. The overall population mean can be expressed in terms of the μ_l as follows. Let x_{il} denote the ith population value in the lth stratum and let $W_l = N_l/N$ denote the fraction of the population in the lth stratum. Then

$$\begin{aligned}\mu &= \frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l} x_{il}\\ &= \frac{1}{N}\sum_{l=1}^{L} N_l\mu_l\\ &= \sum_{l=1}^{L} W_l\mu_l\end{aligned}$$

Within each stratum, a simple random sample of size n_l is taken. The sample mean in stratum l is denoted by

$$\overline{X}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} X_{il}$$

Here X_{il} denotes the ith sample value in the lth stratum. Note that $\overline{X}_l$ is the mean of a simple random sample from the population consisting of the lth stratum, so from Theorem A of Section 7.3.1, $E(\overline{X}_l) = \mu_l$. By analogy with the preceding relationship between the overall population mean and the population means of the various strata, the obvious estimate of μ is

$$\begin{aligned}\overline{X}_s &= \sum_{l=1}^{L} \frac{N_l \overline{X}_l}{N} \\ &= \sum_{l=1}^{L} W_l \overline{X}_l\end{aligned}$$

THEOREM **A**

The stratified estimate, $\overline{X}_s$, of the population mean is unbiased.

Proof

$$\begin{aligned}E(\overline{X}_s) &= \sum_{l=1}^{L} W_l E(\overline{X}_l) \\ &= \frac{1}{N} \sum_{l=1}^{L} N_l \mu_l \\ &= \mu\end{aligned}$$ ■

Since we assume that the samples from different strata are independent of one another and that within each stratum a simple random sample is taken, the variance of $\overline{X}_s$ can be easily calculated.

THEOREM **B**

The variance of the stratified sample mean is given by

$$\text{Var}(\overline{X}_s) = \sum_{l=1}^{L} W_l^2 \left(\frac{1}{n_l}\right) \left(1 - \frac{n_l - 1}{N_l - 1}\right) \sigma_l^2$$

Proof

Since the $\overline{X}_l$ are independent,

$$\text{Var}(\overline{X}_s) = \sum_{l=1}^{L} W_l^2 \text{Var}(\overline{X}_l)$$

From Theorem B of Section 7.3.1, we have

$$\text{Var}(\overline{X}_l) = \frac{1}{n_l} \left(1 - \frac{n_l - 1}{N_l - 1}\right) \sigma_l^2$$

Therefore, the desired result follows. ■

If the sampling fractions within all strata are small,

$$\text{Var}(\overline{X}_s) \approx \sum_{l=1}^{L} \frac{W_l^2 \sigma_l^2}{n_l}$$

EXAMPLE **A** We again consider the population of hospitals. As we did in the discussion of ratio estimates, we assume that the number of beds in each hospital is known but that the number of discharges is not. We will try to make use of this knowledge by stratifying the hospitals according to the number of beds. Let stratum A consist of the 98 smallest hospitals, stratum B of the 98 next larger, stratum C of the 98 next larger, and stratum D of the 99 largest. The following table shows the results of this stratification of hospitals by size:

Stratum	N_l	W_l	μ_l	σ_l
A	98	.249	182.9	103.4
B	98	.249	526.5	204.8
C	98	.249	956.3	243.5
D	99	.251	1591.2	419.2

Suppose that we use a sample of total size n and let

$$n_1 = n_2 = n_3 = n_4 = \frac{n}{4}$$

so that we have equal sample sizes in each stratum. Then, from Theorem B, neglecting the finite population corrections and using the numerical values in the preceding table, we have

$$\begin{aligned}\text{Var}(\overline{X}_s) &= \sum_{l=1}^{4} \frac{W_l^2 \sigma_l^2}{n_1} \\ &= \frac{4}{n} \sum_{l=1}^{4} W_l^2 \sigma_l^2 \\ &= \frac{72,042.6}{n}\end{aligned}$$

and

$$\sigma_{\overline{X}_s} = \frac{268.4}{\sqrt{n}}$$

The standard deviation of the mean of a simple random sample is

$$\sigma_{\overline{X}} = \frac{587.7}{\sqrt{n}}$$

Comparing the two standard deviations, we see that a tremendous gain in precision has resulted from the stratification. The ratio of the variances is .20; thus a stratified estimate based on a total sample size of $n/5$ is as precise as a simple random sample of size n. The reduction in variance due to stratification is comparable to that achieved by using a ratio estimate (Example D in Section 7.4). In later parts of this section, we will look more analytically at why the stratification done here produced such dramatic improvement. ■

Let us next consider the stratified estimate of the population total, $T_s = N\overline{X}_s$. From Theorem B, we have the following corollary.

COROLLARY **A**

The expectation and variance of the stratified estimate of the population total are

$$E(T_s) = \tau$$

and

$$\begin{aligned} \mathrm{Var}(T_s) &= N^2 \mathrm{Var}(\overline{X}_s) \\ &= \sum_{l=1}^{L} N_l^2 \left(\frac{1}{n_l}\right)\left(1 - \frac{n_l - 1}{N_l - 1}\right)\sigma_l^2 \end{aligned}$$ ■

In order to estimate the standard errors of $\overline{X}_s$ and T_s, the variances of the individual strata must be separately estimated and substituted into the preceding formulae. The estimate of σ_l^2 is given by

$$s_l^2 = \frac{1}{n_l - 1}\sum_{i=1}^{n_l}(X_{il} - \overline{X}_l)^2$$

$\mathrm{Var}(\overline{X}_s)$ is estimated by

$$s_{\overline{X}_s}^2 = \sum_{l=1}^{L} W_l^2 \left(\frac{1}{n_l}\right)\left(1 - \frac{n_l}{N_l}\right) s_l^2$$

The next example illustrates how this variance estimate can be used to find approximate confidence intervals for μ based on $\overline{X}_s$.

EXAMPLE **B** A sample of size 10 was drawn from each of the four strata of hospitals described in Example A, yielding the following:

$$\begin{aligned} \overline{X}_1 &= 240.6 & s_1^2 &= 6827.6 \\ \overline{X}_2 &= 507.4 & s_2^2 &= 23{,}790.7 \\ \overline{X}_3 &= 865.1 & s_3^2 &= 42{,}573.0 \\ \overline{X}_4 &= 1716.5 & s_4^2 &= 152{,}099.6 \end{aligned}$$

Therefore, $\overline{X}_s = 832.5$. The variance of the stratified sample mean is estimated by

$$\begin{aligned} s^2_{\overline{X}_s} &= \frac{1}{10}\sum_{l=1}^{4} W_l^2\left(1 - \frac{n_l - 1}{N_l - 1}\right) s_l^2 \\ &= 1282.0 \end{aligned}$$

Thus,

$$s_{\overline{X}_s} = 35.8$$

An approximate 95% confidence interval for the population mean number of discharges is $\overline{X}_s \pm 1.96 s_{\bar{X}_s}$, or (762.4, 902.7).

The total number of discharges is estimated by $T_s = 393\overline{X}_s = 327{,}172$. The standard error of T_s is estimated by $s_{T_s} = 393 s_{\overline{X}_s} = 14{,}069$. An approximate 95% confidence interval for the population total is $T_s \pm 1.96 s_{T_s}$, or (299,596, 354, 748). ■

7.5.3 Methods of Allocation

In Section 7.5.2, it was shown that, neglecting the finite population correction,

$$\text{Var}(\overline{X}_s) = \sum_{l=1}^{L} \frac{W_l^2 \sigma_l^2}{n_l}$$

If the resources of a survey allow only a total of n units to be sampled, the question arises of how to choose $n_1, \ldots, n_L$ to minimize $\text{Var}(\overline{X}_s)$ subject to the constraint $n_1 + \cdots + n_L = n$.

For the sake of simplicity, the calculations in this section ignore the finite population correction within each stratum. The analysis may be extended to include these corrections, but at the cost of some additional algebra. More complete results are contained in Cochran (1977).

THEOREM **A**

The sample sizes $n_1, \ldots, n_L$ that minimize $\text{Var}(\overline{X}_s)$ subject to the constraint $n_1 + \cdots + n_L = n$ are given by

$$n_l = n\frac{W_l \sigma_l}{\sum\limits_{k=1}^{L} W_k \sigma_k}$$

where $l = 1, \ldots, L$.

Proof

We introduce a Lagrange multiplier, and we must then minimize

$$L(n_1, \ldots, n_L, \lambda) = \sum_{l=1}^{L} \frac{W_l^2 \sigma_l^2}{n_l} + \lambda \left(\sum_{l=1}^{L} n_l - n \right)$$

For $l = 1, \ldots, L$, we have

$$\frac{\partial L}{\partial n_l} = -\frac{W_l^2 \sigma_l^2}{n_l^2} + \lambda$$

Setting these partial derivatives equal to zero, we have the system of equations

$$n_l = \frac{W_l \sigma_l}{\sqrt{\lambda}}$$

for $l = 1, \ldots, L$. To determine λ, we first sum these equations over l:

$$n = \frac{1}{\sqrt{\lambda}} \sum_{l=1}^{L} W_l \sigma_l$$

Thus,

$$\frac{1}{\sqrt{\lambda}} = \frac{n}{\sum\limits_{l=1}^{L} W_l \sigma_l}$$

and

$$n_l = n \frac{W_l \sigma_l}{\sum\limits_{l=1}^{L} W_l \sigma_l}$$

which proves the theorem. ■

This theorem shows that those strata for which $W_l \sigma_l$ is large should be sampled heavily. This makes sense intuitively. If W_l is large, the stratum contains a large fraction of the population; if σ_l is large, the population values in the stratum are quite variable, and in order to obtain a good determination of the stratum's mean, a relatively large sample size must be used. This optimal allocation scheme is called **Neyman allocation.**

Substituting the optimal values of n_l as given in Theorem A into the equation for $\mathrm{Var}(\overline{X}_s)$ given in Theorem B in Section 7.5.2 gives us the following corollary.

COROLLARY **A**

Denoting by $\overline{X}_{so}$, the stratified estimate using the optimal allocations as given in Theorem A and neglecting the finite population correction,

$$\mathrm{Var}(\overline{X}_{so}) = \frac{\left(\sum\limits_{l=1}^{L} W_l \sigma_l \right)^2}{n}$$ ■

EXAMPLE **A** For the population of hospitals, the weights for optimal allocation, $W_l\sigma_l / \sum W_l\sigma_l$, are, from the table of Example A of Section 7.5.2,

	Stratum			
	A	B	C	D
Weight	.106	.210	.250	.434

Note that, because of its larger standard deviation, stratum D is sampled more than four times as heavily as stratum A. ■

The optimal allocations depend on the individual variances of the strata, which generally will not be known. Furthermore, if a survey measures several attributes for each population member, it is usually impossible to find an allocation that is simultaneously optimal for all of those variables. A simple and popular alternative method of allocation is to use the same sampling fraction in each stratum,

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_L}{N_L}$$

which holds if

$$n_l = n\frac{N_l}{N} = nW_l$$

for $l = 1, \ldots, L$. This method is called **proportional allocation.** The estimate of the population mean based on proportional allocation is

$$\begin{aligned}\overline{X}_{sp} &= \sum_{l=1}^{L} W_l \overline{X}_l \\ &= \sum_{l=1}^{L} W_l \frac{1}{n_l}\sum_{i=1}^{n_l} X_{il} \\ &= \frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n_l} X_{il}\end{aligned}$$

since $W_l/n_l = 1/n$. This estimate is simply the unweighted mean of the sample values.

THEOREM **B**

With stratified sampling based on proportional allocation, ignoring the finite population correction,

$$\mathrm{Var}(\overline{X}_{sp}) = \frac{1}{n}\sum_{l=1}^{L} W_l\sigma_l^2$$

Proof

From Theorem B of Section 7.5.2, we have

$$\begin{aligned}\mathrm{Var}(\overline{X}_{sp}) &= \sum_{l=1}^{L} W_l^2 \mathrm{Var}(\overline{X}_l)\\ &= \sum_{l=1}^{L} W_l^2 \frac{\sigma_l^2}{n_l}\end{aligned}$$

Using $n_l = nW_l$, the result follows. ■

We now compare $\mathrm{Var}(\overline{X}_{sp})$ and $\mathrm{Var}(\overline{X}_{so})$ in order to discover the circumstances under which optimal allocation is substantially better than proportional allocation.

THEOREM **C**

With stratified random sampling, the difference between the variance of the estimate of the population mean based on proportional allocation and the variance of that estimate based on optimal allocation is, ignoring the finite population correction,

$$\mathrm{Var}(\overline{X}_{sp}) - \mathrm{Var}(\overline{X}_{so}) = \frac{1}{n}\sum_{l=1}^{L} W_l(\sigma_l - \bar{\sigma})^2$$

where

$$\bar{\sigma} = \sum_{l=1}^{L} W_l \sigma_l$$

Proof

$$\mathrm{Var}(\overline{X}_{sp}) - \mathrm{Var}(\overline{X}_{so}) = \frac{1}{n}\left[\sum_{l=1}^{L} W_l\sigma_l^2 - \left(\sum_{l=1}^{L} W_l\sigma_l\right)^2\right]$$

The term within the large brackets equals $\sum_{l=1}^{L} W_l(\sigma_l - \bar{\sigma})^2$, which may be verified by expanding the square and collecting terms. ■

According to Theorem C, if the variances of the strata are all the same, proportional allocation yields the same results as optimal allocation. The more variable these variances are, the better it is to use optimal allocation.

EXAMPLE **B** Let us calculate how much better optimal allocation is than proportional allocation for the population of hospitals. From Theorem C and Corollary A, we have

$$\operatorname{Var}(\overline{X}_{sp}) = \operatorname{Var}(\overline{X}_{so}) + \frac{1}{n}\sum W_l(\sigma_l - \bar{\sigma})^2$$

Therefore,

$$\begin{aligned}\frac{\operatorname{Var}(\overline{X}_{sp})}{\operatorname{Var}(\overline{X}_{so})} &= 1 + \frac{\dfrac{1}{n}\sum W_l(\sigma_l - \bar{\sigma})^2}{\operatorname{Var}(\overline{X}_{so})}\\ &= 1 + \frac{\sum W_l(\sigma_l - \bar{\sigma})^2}{(\sum W_l\sigma_l)^2}\\ &= 1 + .218\end{aligned}$$

Thus, under proportional allocation, the variance of the mean is about 20% larger than it is under optimal allocation. ■

We can also compare the variance under simple random sampling with the variance under proportional allocation. The variance under simple random sampling is, neglecting the finite population correction,

$$\operatorname{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

In order to compare this equation with that for the variance under proportional allocation, we need a relationship between the overall population variance, σ^2, and the strata variances, σ_l^2. The overall population variance may be expressed as

$$\sigma^2 = \frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l}(x_{il} - \mu)^2$$

Also,

$$\begin{aligned}(x_{il} - \mu)^2 &= [(x_{il} - \mu_l) + (\mu_l - \mu)]^2\\ &= (x_{il} - \mu_l)^2 + 2(x_{il} - \mu_l)(\mu_l - \mu) + (\mu_l - \mu)^2\end{aligned}$$

When both sides of this last equation are summed over l, the middle term on the right-hand side becomes zero since $N_l\mu_l = \sum_{l=1}^{N_l} x_{il}$, so we have

$$\begin{aligned}\sum_{i=1}^{N_l}(x_{il} - \mu)^2 &= \sum_{i=1}^{n_l}(x_{il} - \mu_l)^2 + N_l(\mu_l - \mu)^2\\ &= N_l\sigma_l^2 + N_l(\mu_l - \mu)^2\end{aligned}$$

Dividing both sides by N and summing over l, we have

$$\sigma^2 = \sum_{l=1}^{L} W_l\sigma_l^2 + \sum_{l=1}^{L} W_l(\mu_l - \mu)^2$$

Substituting this expression for σ^2 into $\text{Var}(\overline{X}) = \sigma^2/n$ and using the formula for $\text{Var}(\overline{X}_{sp})$ given in Theorem B completes a proof of the following theorem.

THEOREM **D**

The difference between the variance of the mean of a simple random sample and the variance of the mean of a stratified random sample based on proportional allocation is, neglecting the finite population correction,

$$\text{Var}(\overline{X}) - \text{Var}(\overline{X}_{sp}) = \frac{1}{n}\sum_{l=1}^{L} W_l(\mu_l - \mu)^2 \qquad \blacksquare$$

Thus, stratified random sampling with proportional allocation always gives a smaller variance than does simple random sampling, providing that the finite population correction is ignored. Comparing the equations for the variances under simple random sampling, proportional allocation, and optimal allocation, we see that stratification with proportional allocation is better than simple random sampling if the strata means are quite variable and that stratification with optimal allocation is even better than stratification with proportional allocation if the strata standard deviations are variable.

EXAMPLE **C** We calculate the improvement that would result from using stratification with proportional allocation rather than simple random sampling for the population of hospitals. From Theorems B and D, we have

$$\begin{aligned}\frac{\text{Var}(\overline{X}_{srs})}{\text{Var}(\overline{X}_{sp})} &= 1 + \frac{\sum W_l(\mu_l - \bar{\mu})^2}{\sum W_l\sigma_l^2} \\ &= 1 + 3.83\end{aligned}$$

As is frequently the case, the gain from using stratification with proportional allocation rather than simple random sampling is much greater than the gain from using optimal allocation rather than proportional allocation. Furthermore, proportional allocation requires knowledge only of the sizes of the strata, whereas optimal allocation requires knowledge of the standard deviations of the strata, and such knowledge is usually unavailable. ■

Typically, stratified random sampling can result in substantial increases in precision for populations containing values that vary greatly in size. For example, a population of transactions, a sample of which is to be audited for errors, might contain transactions in the hundreds of thousands of dollars and transactions in the hundreds of dollars. If such a population were divided into several strata according to the dollar amounts of the transactions, there might well be considerable variation in the mean transaction errors between the strata, since there may be rather large errors on large transactions and small errors on small transactions. The variability of the errors might also be larger in the former strata as well.

We have not addressed the question of how many strata to form and how to define the strata. In order to construct the optimal number of strata, the population values themselves, which are of course unknown, would have to be used. Stratification must therefore be done on the basis of some related variable that is known (such as transaction amount in the preceding paragraph) or on the results of earlier samples. In practice, it usually turns out that such relationships are not strong enough to make it worthwhile constructing more than a few strata.

7.6 Concluding Remarks

This chapter introduced survey sampling. It first covered the most elementary method of probability sampling—simple random sampling. The theory of this method underlies the theory of more complex sampling techniques. Stratified sampling was also introduced and shown to increase the precision of estimates substantially in many cases.

Several concepts and techniques introduced here recur throughout statistics: the concept of a random estimate of a population parameter, such as the population mean; bias; the standard error of an estimate; confidence intervals based on the central limit theorem; and linearization, or propagation of error.

The theory and technique of survey sampling go far beyond the material in this introduction. One method that deserves mention because of its widespread use is **systematic sampling.** The population members are given in a list. If, say, a 10% sample is desired, every tenth member of the list is sampled starting from some random point among the first ten. If the list is in totally random order, this method is similar to simple random sampling. If, however, there is some correlation or relationship between successive members, the method is more similar to stratified sampling. The clear danger of this method is that there may be some periodic structure in the list, in which case bias can ensue.

Another commonly used method is **cluster sampling.** In sampling residential households, a survey might choose blocks randomly and then either sample every dwelling on each chosen block or further subsample the dwellings. Because one would expect dwellings within a single block to be relatively homogeneous, this method is typically less precise than a simple random sample of the same size.

We have developed a mathematical model for survey sampling and have deduced consequences of that model, including probabilistic error bounds for the estimates. As is always the case, reality never quite matches the mathematical model. The basic assumptions of the model are (1) that every population member appears in the sample with a specified probability and (2) that an exact measurement or response is obtained from every sample member. In practice, neither assumption will hold precisely. Converse and Traugott (1986) provide an interesting discussion of the practical difficulties of polls and surveys and consequences for the variability of the estimates.

The first assumption may fail because of the difficulty of obtaining an exact enumeration of the population or because of imprecision in its definition. For example, political surveys can be putatively based on all adults, all registered voters, or all “likely” voters. However, the most serious problem with respect to the first assumption is that of nonresponse. Response levels of only 60% to 70% are common in surveys of human populations. The possibility of substantial bias clearly arises if there is a relationship of potential answers to survey questions to the propensity to respond to those questions. For example, adults living in families are easier to contact by a telephone survey than those living alone, and the opinions of these two groups may well differ on certain issues. It is important to realize that the standard errors of estimates that we have developed earlier in this chapter account only for random variability in sample composition, not for systematic biases.

The *Literary Digest* poll of 1936, which predicted a 57% to 43% victory for Republican Alfred Landon over incumbent president Franklin Roosevelt, is one of the most famous of flawed surveys. Questionnaires were mailed to about 10 million voters, who were selected from lists such as telephone books and club memberships, and approximately 2.4 million of the questionnaires were returned. There were two intrinsic problems: (1) nonresponse—those who did not respond may have voted differently from those who did—and (2) selection bias—even if all 10 million voters had responded, they would not have constituted a random sample; those in lower socioeconomic classes (who were more likely to vote for Roosevelt) were less likely to have telephone service or belong to clubs and thus less likely to be included in the sample than were wealthier voters. The assumption that an exact measurement is obtained from every member of the sample may also be in error. In surveys conducted by interviewers, the interviewer's approach and personality may affect the response. In surveys that use questionnaires, the wording of the questions and the context within which they are lodged can have an effect. An interesting example is a poll conducted by Stanley Presser, (*New Yorker,* Oct 18, 2004). Half of the sample was asked, "Do you think the United States should allow public speeches against democracy?" The other half was asked, "Do you think the United States should forbid public speeches against democracy?" 56% said no to the first question, and 39% said yes to the second. The interesting paper by Hansen in Tanur et al. (1972) reports on efforts of the U.S. Bureau of the Census to investigate these sorts of problems.

7.7 Problems

1. Consider a population consisting of five values—1, 2, 2, 4, and 8. Find the population mean and variance. Calculate the sampling distribution of the mean of a sample of size 2 by generating all possible such samples. From them, find the mean and variance of the sampling distribution, and compare the results to Theorems A and B in Section 7.3.1.

2. Suppose that a sample of size $n = 2$ is drawn from the population of the preceding problem and that the proportion of the sample values that are greater than 3 is recorded. Find the sampling distribution of this statistic by listing all possible such samples. Find the mean and variance of the sampling distribution.

3. Which of the following is a random variable?

a. The population mean
b. The population size, N

c. The sample size, n
d. The sample mean
e. The variance of the sample mean
f. The largest value in the sample
g. The population variance
h. The estimated variance of the sample mean

4. Two populations are surveyed with simple random samples. A sample of size n_1 is used for population I, which has a population standard deviation σ_1; a sample of size $n_2 = 2n_1$ is used for population II, which has a population standard deviation $\sigma_2 = 2\sigma_1$. Ignoring finite population corrections, in which of the two samples would you expect the estimate of the population mean to be more accurate?

5. How would you respond to a friend who asks you, "How can we say that the sample mean is a random variable when it is just a number, like the population mean? For example, in Example A of Section 7.3.2, a simple random sample of size 50 produced $\bar{x} = 938.5$; how can the number 938.5 be a random variable?"

6. Suppose that two populations have equal population variances but are of different sizes: $N_1 = 100{,}000$ and $N_2 = 10{,}000{,}000$. Compare the variances of the sample means for a sample of size $n = 25$. Is it substantially easier to estimate the mean of the smaller population?

7. Suppose that a simple random sample is used to estimate the proportion of families in a certain area that are living below the poverty level. If this proportion is roughly .15, what sample size is necessary so that the standard error of the estimate is .02?

8. A sample of size $n = 100$ is taken from a population that has a proportion $p = 1/5$.

a. Find δ such that $P(|\hat{p} - p| \geq \delta) = 0.025$.
b. If, in the sample, $\hat{p} = 0.25$, will the 95% confidence interval for p contain the true value of p?

9. In a simple random sample of 1,500 voters, 55% said they planned to vote for a particular proposition, and 45% said they planned to vote against it. The estimated margin of victory for the proposition is thus 10%. What is the standard error of this estimated margin? What is an approximate 95% confidence interval for the margin?

10. True or false (and state why):
If a sample from a population is large, a histogram of the values in the sample will be approximately normal, even if the population is not normal.

11. Consider a population of size four, the members of which have values x_1, x_2, x_3, x_4.

a. If simple random sampling were used, how many samples of size two are there?
b. Suppose that rather than simple random sampling, the following sampling scheme is used. The possible samples of size two are

$$\{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_4\}, \{x_1, x_4\}$$

and the sampling is done in such a way that each of these four possible samples is equally likely. Is the sample mean unbiased?

12. Consider simple random sampling *with* replacement.

a. Show that

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimate of σ^2.

b. Is s an unbiased estimate of σ?

c. Show that $n^{-1}s^2$ is an unbiased estimate of $\sigma^2_{\overline{X}}$.

d. Show that $n^{-1}N^2s^2$ is an unbiased estimate of σ^2_T.

e. Show that $\hat{p}(1-\hat{p})/(n-1)$ is an unbiased estimate of $\sigma^2_{\hat{p}}$.

13. Suppose that the total number of discharges, τ, in Example A of Section 7.2 is estimated from a simple random sample of size 50. Denoting the estimate by T, use the central limit theorem to sketch the approximate probability density of the error $T - \tau$.

14. The proportion of hospitals in Example A of Section 7.2 that had fewer than 1000 discharges is $p = .654$. Suppose that the total number of hospitals having fewer than 1000 discharges is estimated from a simple random sample of size 25. Use the central limit theorem to sketch the approximate sampling distribution of the estimate.

15. Consider estimating the mean of the population of hospital discharges (Example A of Section 7.2) from a simple random sample of size n. Use the normal approximation to the distribution of $\overline{X}$ in answering the following:

a. Sketch $P(|\overline{X} - \mu| > 200)$ as a function of n for $20 \le n \le 100$.

b. For $n = 20, 40,$ and 80, find Δ such that $P(|\overline{X} - \mu| > \Delta) \approx .10$. Similarly, find Δ such that $P(|\overline{X} - \mu| > \Delta) \approx .50$.

16. True or false?

a. The center of a 95% confidence interval for the population mean is a random variable.

b. A 95% confidence interval for μ contains the sample mean with probability .95.

c. A 95% confidence interval contains 95% of the population.

d. Out of one hundred 95% confidence intervals for μ, 95 will contain μ.

17. A 90% confidence interval for the average number of children per household based on a simple random sample is found to be (.7, 2.1). Can we conclude that 90% of households have between .7 and 2.1 children?

18. From independent surveys of two populations, 90% confidence intervals for the population means are constructed. What is the probability that neither interval contains the respective population mean? That both do?

19. This problem introduces the concept of a *one-sided confidence interval.* Using the central limit theorem, how should the constant k be chosen so that the interval

$(-\infty, \overline{X} + ks_{\overline{X}})$ is a 90% confidence interval for μ—i.e., so that $P(\mu \leq \overline{X} + ks_{\overline{X}}) = .9$? This is called a one-sided confidence interval. How should k be chosen so that $(\overline{X} - ks_{\overline{X}}, \infty)$ is 95% one-sided confidence interval?

20. In Example D of Section 7.3.3, a 95% confidence interval for μ was found to be $(1.44, 1.76)$. Because μ is some fixed number, it either lies in this interval or it doesn't, so it doesn't make any sense to claim that $P(1.44 \leq \mu \leq 1.76) = .95$. What do we mean, then, by saying this is a "95% confidence interval?"

21. In order to halve the width of a 95% confidence interval for a mean, by what factor should the sample size be increased? Ignore the finite population correction.

22. An investigator quantifies her uncertainty about the estimate of a population mean by reporting $\overline{X} \pm s_{\overline{X}}$. What size confidence interval is this?

23. **a.** Show that the standard error of an estimated proportion is largest when $p = 1/2$.

b. Use this result and Corollary B of Section 7.3.2 to conclude that the quantity

$$\frac{1}{2}\sqrt{\frac{N-n}{N(n-1)}}$$

is a conservative estimate of the standard error of $\hat{p}$ no matter what the value of p may be.

c. Use the central limit theorem to conclude that the interval

$$\hat{p} \pm \sqrt{\frac{N-n}{N(n-1)}}$$

contains p with probability at least .95.

24. For a random sample of size n from a population of size N, consider the following as an estimate of μ:

$$\overline{X}_c = \sum_{i=1}^{n} c_i X_i$$

where the c_i are fixed numbers and $X_1, \ldots, X_n$ is the sample.

a. Find a condition on the c_i such that the estimate is unbiased.

b. Show that the choice of c_i that minimizes the variances of the estimate subject to this condition is $c_i = 1/n$, where $i = 1, \ldots, n$.

25. Here is an alternative proof of Lemma B in Section 7.3.1. Consider a random permutation $Y_1, Y_2, \ldots, Y_N$ of $x_1, x_2, \ldots, x_N$. Argue that the joint distribution of any subcollection, $Y_{i_1}, \ldots, Y_{i_n}$, of the Y_i is the same as that of a simple random sample, $X_1, \ldots, X_n$. In particular,

$$\text{Var}(Y_i) = \text{Var}(X_k) = \sigma^2$$

and

$$\text{Cov}(Y_i, Y_j) = \text{Cov}(X_k, X_l) = \gamma$$

if $i \neq j$ and $k \neq l$. Since $Y_1 + Y_2 + \cdots + Y_N = \tau$,

$$\operatorname{Var}\left(\sum_{i=1}^{N} Y_i\right) = 0$$

(Why?) Express $\operatorname{Var}(\sum_{i=1}^{N} Y_i)$ in terms of σ^2 and the unknown covariance, γ. Solve for γ, and conclude that

$$\gamma = -\frac{\sigma^2}{N-1}$$

for $i \neq j$.

26. This is another proof of Lemma B in Section 7.3.1. Let U_i be a random variable with $U_i = 1$ if the ith population member is in the sample and equal to 0 otherwise.

a. Show that the sample mean $\overline{X} = n^{-1} \sum_{i=1}^{N} U_i x_i$.

b. Show that $P(U_i = 1) = n/N$. Find $E(U_i)$, using the fact that U_i is a Bernoulli random variable.

c. What is the variance of the Bernoulli random variable U_i?

d. Noting that $U_i U_j$ is a Bernoulli random variable, find $E(U_i U_j)$, $i \neq j$. (Be careful to take into account that the sample is drawn without replacement.)

e. Find $\operatorname{Cov}(U_i, U_j)$, $i \neq j$.

f. Using the representation of $\overline{X}$ above, find $\operatorname{Var}(\overline{X})$.

27. Suppose that the population size N is not known, but it is known that $n \leq N$. Show that the following procedure will generate a simple random sample of size n. Imagine that the population is arranged in a long list that you can read sequentially.

a. Let the sample initially consist of the the first n elements in the list.

b. For $k = 1, 2, \ldots,$ as long as the end of the list has not been encountered:

i. Read the $(n + k)$-th element in the list.

ii. Place it in the sample with probability $n/(n + k)$ and, if it is placed in the sample, randomly drop one of the exisiting sample members.

28. In surveys, it is difficult to obtain accurate answers to sensitive questions such as "Have you ever used heroin?" or "Have you ever cheated on an exam?" Warner (1965) introduced the method of **randomized response** to deal with such situations. A respondent spins an arrow on a wheel or draws a ball from an urn containing balls of two colors to determine which of two statements to respond to: (1) "I have characteristic A," or (2) "I do not have characteristic A." The interviewer does not know which statement is being responded to but merely records a yes or a no. The hope is that an interviewee is more likely to answer truthfully if he or she realizes that the interviewer does not know which statement is being responded to. Let R be the proportion of a sample answering Yes. Let p be the probability that statement 1 is responded to (p is known from the structure of the randomizing device), and let q be the proportion of the population that has characteristic A. Let r be the probability that a respondent answers Yes.

a. Show that $r = (2p-1)q + (1-p)$. [*Hint:* $P(\text{yes}) = P(\text{yes given question 1}) \times P(\text{question 1}) + P(\text{yes given question 2}) \times P(\text{question 2})$.]

b. If r were known, how could q be determined?

c. Show that $E(R) = r$, and propose an estimate, Q, for q. Show that the estimate is unbiased.

d. Ignoring the finite population correction, show that

$$\mathrm{Var}(R) = \frac{r(1-r)}{n}$$

where n is the sample size.

e. Find an expression for $\mathrm{Var}(Q)$.

29. A variation of the method described in Problem 28 has been proposed. Instead of responding to statement 2, the respondent answers an unrelated question for which the probability of a "yes" response is known, for example, "Were you born in June?"

a. Propose an estimate of q for this method.

b. Show that the estimate is unbiased.

c. Obtain an expression for the variance of the estimate.

30. Compare the accuracies of the methods of Problems 28 and 29 by comparing their standard deviations. You may do this by substituting some plausible numerical values for p and q.

31. Referring to Example D in Section 7.3.3, how large should the sample be in order that the 95% confidence interval for the total number of owners planning to sell will have a width of 500?

32. Referring again to Example D in Section 7.3.3, suppose that a survey is done of another condominium project of 12,000 units. The sample size is 200, and the proportion planning to sell in this sample is .18.

a. What is the standard error of this estimate? Give a 90% confidence interval.

b. Suppose we use the notation $\hat{p}_1 = .12$ and $\hat{p}_2 = .18$ to refer to the proportions in the two samples. Let $\hat{d} = \hat{p}_1 - \hat{p}_2$ be an estimate of the difference, d, of the two population proportions p_1 and p_2. Using the fact that $\hat{p}_1$ and $\hat{p}_2$ are independent random variables, find expressions for the variance and standard error of $\hat{d}$.

c. Because $\hat{p}_1$ and $\hat{p}_2$ are approximately normally distributed, so is $\hat{d}$. Use this fact to construct 99%, 95%, and 90% confidence intervals for d. Is there clear evidence that p_1 is really different from p_2?

33. Two populations are independently surveyed using simple random samples of size n, and two proportions, p_1 and p_2, are estimated. It is expected that both population proportions are close to .5. What should the sample size be so that the standard error of the difference, $\hat{p}_1 - \hat{p}_2$, will be less than .02?

34. In a survey of a very large population, the incidences of two health problems are to be estimated from the same sample. It is expected that the first problem will affect about 3% of the population and the second about 40%. Ignore the finite population correction in answering the following questions.

a. How large should the sample be in order for the standard errors of both estimates to be less than .01? What are the actual standard errors for this sample size?

b. Suppose that instead of imposing the same limit on both standard errors, the investigator wants the standard error to be less than 10% of the true value in each case. What should the sample size be?

35. A simple random sample of a population of size 2000 yields the following 25 values:

104	109	111	109	87
86	80	119	88	122
91	103	99	108	96
104	98	98	83	107
79	87	94	92	97

a. Calculate an unbiased estimate of the population mean.

b. Calculate unbiased estimates of the population variance and $\mathrm{Var}(\overline{X})$.

c. Give approximate 95% confidence intervals for the population mean and total.

36. With simple random sampling, is $\overline{X}^2$ an unbiased estimate of μ^2? If not, what is the bias?

37. Two surveys were independently conducted to estimate a population mean, μ. Denote the estimates and their standard errors by $\overline{X}_1$ and $\overline{X}_2$ and $\sigma_{\overline{X}_1}$ and $\sigma_{\overline{X}_2}$. Assume that $\overline{X}_1$ and $\overline{X}_2$ are unbiased. For some α and β, the two estimates can be combined to give a better estimator:

$$X = \alpha \overline{X}_1 + \beta \overline{X}_2$$

a. Find the conditions on α and β that make the combined estimate unbiased.

b. What choice of α and β minimizes the variances, subject to the condition of unbiasedness?

38. Let $X_1, \ldots, X_n$ be a simple random sample. Show that $\frac{1}{n}\sum_{i=1}^{n} X_i^3$ is an unbiased estimate of $\frac{1}{N}\sum_{i=1}^{N} x_i^3$.

39. Suppose that of a population of N items, k are defective in some way. For example, the items might be documents, a small proportion of which are fraudulent. How large should a sample be so that with a specified probability it will contain at least one of the defective items? For example, if $N = 10{,}000$, $k = 50$, and $p = .95$, what should the sample size be? Such calculations are useful in planning sample sizes for acceptance sampling.

40. This problem presents an algorithm for drawing a simple random sample from a population in a sequential manner. The members of the population are considered for inclusion in the sample one at a time in some prespecified order (for example, the order in which they are listed). The ith member of the population is included in the sample with probability

$$\frac{n - n_i}{N - i + 1}$$

where n_i is the number of population members already in the sample before the ith member is examined. Show that the sample selected in this way is in fact a simple random sample; that is, show that every possible sample occurs with probability

$$\frac{1}{\binom{N}{n}}$$

41. In accounting and auditing, the following sampling method is sometimes used to estimate a population total. In estimating the value of an inventory, suppose that a book value exists for each item and is readily accessible. For each item in the sample, the difference D, audited value minus book value, is determined. The inventory value is estimated by the sum of the book values of the population and $N\overline{D}$, where N is the population size.

a. Show that the estimate is unbiased.
b. Find an expression for the variance of the estimate.
c. Compare the expression obtained in part (b) to the variance of the usual estimate, which is the product of N and the average audited value. Under what circumstances would the proposed method be more accurate?
d. How could a ratio estimate be employed in this situation? Would there be any advantage or disadvantage to using a ratio estimate rather than the proposed method?

42. Show that the population correlation coefficient is less than or equal to 1 in absolute value.

43. Suppose that for Example D in Section 7.3.3, the average number of occupants per condominium unit in the sample is 2.2 with a sample standard deviation of .7 and the sample correlation coefficient between the number of occupants and the number of motor vehicles is .85. Estimate the population ratio of the number of motor vehicles per occupant and its standard error. Find an approximate 95% confidence interval for the estimate.

44. Show that

$$\frac{\mathrm{Var}(\overline{Y}_R)}{\mathrm{Var}(\overline{Y})} \approx 1 + \frac{C_x}{C_y}\left(\frac{C_x}{C_y} - 2\rho\right)$$

Sketch the graph of this ratio as a function of C_x/C_y.

45. In the population of hospitals, the correlation of the number of beds and the number of discharges is $\rho = .91$ (Example D of Section 7.4). To see how $\mathrm{Var}(\overline{Y}_R)$ would be different if the correlation were different, plot $\mathrm{Var}(\overline{Y}_R)$ for $n = 64$ as a function of ρ for $-1 < \rho < 1$.

46. Use the central limit theorem to sketch the approximate sampling distribution of $\overline{Y}_R$ for $n = 64$ for the population of hospitals. Compare to the approximate sampling distribution of $\overline{Y}$.

47. For the population of hospitals and a sample size of $n = 64$, find the approximate bias of $\overline{Y}_R$ by applying Corollary B of Section 7.4 and compare it to the approximate standard deviation of the estimate. Repeat for $n = 128$.

48. A simple random sample of 100 households located in a city recorded the number of people living in the household, X, and the weekly expenditure for food, Y. It is known that there are 100,000 households in the city. In the sample

$$\sum X_i = 320$$
$$\sum Y_i = 10{,}000$$
$$\sum X_i^2 = 1250$$
$$\sum Y_i^2 = 1{,}100{,}000$$
$$\sum X_i Y_i = 36{,}000$$

Neglect the finite population correction in answering the following.

a. Estimate the ratio $r = \mu_y/\mu_x$.
b. Form an approximate 95% confidence interval for μ_y/μ_x.
c. Using only the data on Y estimate the total weekly food expenditure, τ, for households in the city and form a 90% confidence interval.

49. In a wildlife survey, an area of desert land was divided into 1000 squares, or "quadrats," a simple random sample of 50 of which were surveyed. In each surveyed quadrat, the number of birds, Y, and the area covered by vegetation, X, were determined. It was found that

$$\sum X_i = 3000$$
$$\sum Y_i = 150$$
$$\sum X_i^2 = 225{,}000$$
$$\sum Y_i^2 = 650$$
$$\sum X_i Y_i = 11{,}000$$

a. Estimate the ratio of the average number of birds per quadrat to the average vegetation cover per quadrat.
b. Estimate the standard error of your estimate and find an approximate 90% confidence interval for the population average.
c. Estimate the total number of birds and find an approximate 95% confidence interval for the population total.
d. Suppose that from an aerial survey, the total area covered by vegetation could easily be determined. How could this information be used to provide another

estimate of the number of birds? Would you expect this estimate to be better than or worse than that found in part (c)?

50. Hartley and Ross (1954) derived the following exact bound on the relative size of the bias and standard error of a ratio estimate:

$$\frac{|E(R) - r|}{\sigma_R} \leq \frac{\sigma_{\overline{X}}}{\mu_x} = \frac{\sigma_x}{\mu_x}\sqrt{\frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)}$$

a. Derive this bound from the relation

$$\text{Cov}(R, \overline{X}) = E\left(\frac{\overline{Y}}{\overline{X}}\overline{X}\right) - E\left(\frac{\overline{Y}}{\overline{X}}\right)E(\overline{X})$$

b. Apply the bound to Problem 43 using sample estimates in place of the given population parameters.

51. This problem introduces a technique called the "jackknife," originally proposed by Quenouille (1956) for reducing bias. Many nonlinear estimates, including the ratio estimator, have the property that

$$E(\hat{\theta}) = \theta + \frac{b_1}{n} + \frac{b_2}{n^2} + \cdots$$

where $\hat{\theta}$ is an estimate of θ. The jackknife forms an estimate $\hat{\theta}_J$, which has a leading bias term of the order n^{-2} rather than n^{-1}. Thus, for sufficiently large n, the bias of $\hat{\theta}_J$ is substantially smaller than that of $\hat{\theta}$. The technique involves splitting the sample into several subsamples, computing the estimate for each subsample, and then combining the several estimates. The sample is split into p groups of size m, where $n = mp$. For $j = 1, \ldots, p$, the estimate $\hat{\theta}_j$ is calculated from the $m(p-1)$ observations left after the jth group has been deleted. From the preceding expression,

$$E(\hat{\theta}_j) = \theta + \frac{b_1}{m(p-1)} + \frac{b_2}{[m(p-1)]^2} + \cdots$$

Now, p "pseudovalues" are defined:

$$V_j = p\hat{\theta} - (p-1)\hat{\theta}_j$$

The jackknife estimate, $\hat{\theta}_J$, is defined as the average of the pseudovalues:

$$\hat{\theta}_J = \frac{1}{p}\sum_{j=1}^{p} V_j$$

Show that the bias of $\hat{\theta}_J$ is of the order n^{-2}.

52. A population consists of three strata with $N_1 = N_2 = 1000$ and $N_3 = 500$. A stratified random sample with 10 observations in each stratum yields the following data:

Stratum 1	94	99	106	106	101	102	122	104	97	97
Stratum 2	183	183	179	211	178	179	192	192	201	177
Stratum 3	343	302	286	317	289	284	357	288	314	276

Estimate the population mean and total and give a 90% confidence interval.

53. The following table (Cochran 1977) shows the stratification of all farms in a county by farm size and the mean and standard deviation of the number of acres of corn in each stratum.

Farm Size	N_l	μ_l	σ_l
0–40	394	5.4	8.3
41–80	461	16.3	13.3
81–120	391	24.3	15.1
121–160	334	34.5	19.8
161–200	169	42.1	24.5
201–240	113	50.1	26.0
241 +	148	63.8	35.2

a. For a sample size of 100 farms, compute the sample sizes from each stratum for proportional and optimal allocation, and compare them.

b. Calculate the variances of the sample mean for each allocation and compare them to each other and to the variance of an estimate formed from simple random sampling.

c. What are the population mean and variance?

d. Suppose that ten farms are sampled per stratum. What is $\text{Var}(\overline{X}_s)$? How large a simple random sample would have to be taken to attain the same variance? Ignore the finite population correction.

e. Repeat part (d) using proportional allocation of the 70 samples.

54. **a.** Suppose that the cost of a survey is $C = C_0 + C_1 n$, where C_0 is a startup cost and C_1 is the cost per observation. For a given cost C, find the allocation $n_1, \ldots, n_L$ to L strata that is optimal in the sense that it minimizes the variance of the estimate of the population mean subject to the cost constraint.

b. Suppose that the cost of an observation varies from stratum to stratum—in some strata the observations might be relatively cheap and in others relatively expensive. The cost of a survey with an allocation $n_1, \ldots, n_L$ is

$$C = C_0 + \sum_{l=1}^{L} C_l n_l$$

For a fixed total cost C, what choice of $n_1, \cdots, n_L$ minimizes the variance?

c. Assuming that the cost function is as given in part (b), for a fixed variance, find n_l to minimize cost.

55. The designer of a sample survey stratifies a population into two strata, H and L. H contains 100,000 people, and L contains 500,000. He decides to allocate 100 samples to stratum H and 200 to stratum L, taking a simple random sample in each stratum.

a. How should the designer estimate the population mean?
b. Suppose that the population standard deviation in stratum H is 20 and the standard deviation in stratum L is 10. What will be the standard error of his estimate?
c. Would it be better to allocate 200 samples to stratum H and 100 to stratum L?
d. Would it be better to use proportional allocation?

56. How might stratification be used in each of the following sampling problems?

a. A survey of household expenditures in a city.
b. A survey to examine the lead concentration in the soil in a large plot of land.
c. A survey to estimate the number of people who use elevators in a large building with a single bank of elevators.
d. A survey of programs on a television station, taken to estimate the proportion of time taken up by advertising on Monday through Friday from 6 P.M. until 10 P.M. Assume that 52 weeks of recorded broadcasts are available for analysis.

57. Consider stratifying the population of Problem 1 into two strata: (1, 2, 2) and (4, 8). Assuming that one observation is taken from each stratum, find the sampling distribution of the estimate of the population mean and the mean and standard deviation of the sampling distribution. Compare to Theorems A and B in Section 7.5.2 and the results of Problem 1.

58. (Computer Exercise) Construct a population consisting of the integers from 1 to 100. Simulate the sampling distribution of the sample mean of a sample of size 12 by drawing 100 samples of size 12 and making a histogram of the results.

59. (Computer Exercise) Continuing with Problem 58, divide the population into two strata of equal size, allocate six observations per stratum, and simulate the distribution of the stratified estimate of the population mean. Do the same thing with four strata. Compare the results to each other and to the results of Problem 58.

60. A population consists of two strata, H and L, of sizes 100,000 and 500,000 and standard deviations 20 and 12, respectively. A stratified sample of size 100 is to be taken.

a. Find the optimal allocation for estimating the population mean.
b. Find the optimal allocation for estimating the difference of the means of the strata, $\mu_H - \mu_L$.

61. The value of a population mean increases linearly through time: $\mu(t) = \alpha + \beta t$ while the variance remains constant. Independent simple random samples of size n are taken at times $t = 1, 2,$ and 3.

a. Find conditions on w_1, w_2, and w_3 such that

$$\hat{\beta} = w_1 \overline{X}_1 + w_2 \overline{X}_2 + w_3 \overline{X}_3$$

is an unbiased estimate of the rate of change, β. Here $\overline{X}_i$ denotes the sample mean at time t_i.

b. What values of the w_i minimize the variance subject to the constraint that the estimate is unbiased?

62. In Example B of Section 7.5.2, the standard error of $\overline{X}_s$ was estimated to be $s_{\overline{X}_s} = 35.8$. How good is this estimate—what is the actual standard error of $\overline{X}_s$?

63. (Open-ended) Monte Carlo evaluation of an integral was introduced in Example A of Section 5.2. Refer to that example for the following notation. Try to interpret that method from the point of view of survey sampling by considering an "infinite population" of numbers in the interval [0, 1], each population member x having a value $f(x)$. Interpret $\hat{I}(f)$ as the mean of a simple random sample. What is the standard error of $\hat{I}(f)$? How could it be estimated? How could a confidence interval for $I(f)$ be formed? Do you think that anything could be gained by stratifying the "population?" For example, the strata could be the intervals [0, .5) and [.5, 1]. You might find it helpful to consider some examples.

64. The value of an inventory is to be estimated by sampling. The items are stratified by book value in the following way:

Stratum	N_l	μ_l	σ_l
\$1000 +	70	3000	1250
\$200–1000	500	500	100
\$1–200	10,000	90	30

a. What should the relative sampling fraction in each stratum be for proportional and for optimal allocation? Ignore the finite population correction.

b. How do the variances under each type of allocation compare to each other and to the variance under simple random sampling?

65. The disk file `cancer` contains values for breast cancer mortality from 1950 to 1960 (y) and the adult white female population in 1960 (x) for 301 counties in North Carolina, South Carolina, and Georgia.

a. Make a histogram of the population values for cancer mortality.

b. What are the population mean and total cancer mortality? What are the population variance and standard deviation?

c. Simulate the sampling distribution of the mean of a sample of 25 observations of cancer mortality.

d. Draw a simple random sample of size 25 and use it to estimate the mean and total cancer mortality.

e. Estimate the population variance and standard deviation from the sample of part (d).

f. Form 95% confidence intervals for the population mean and total from the sample of part (d). Do the intervals cover the population values?

g. Repeat parts (d) through (f) for a sample of size 100.

h. Suppose that the size of the total population of each county is known and that this information is used to improve the cancer mortality estimates by forming a ratio estimator. Do you think this will be effective? Why or why not?

i. Simulate the sampling distribution of ratio estimators of mean cancer mortality based on a simple random sample of size 25. Compare this result to that of part (c).

j. Draw a simple random sample of size 25 and estimate the population mean and total cancer mortality by calculating ratio estimates. How do these estimates compare to those formed in the usual way in part (d) from the same data?

k. Form confidence intervals about the estimates obtained in part (j).

l. Stratify the counties into four strata by population size. Randomly sample six observations from each stratum and form estimates of the population mean and total mortality.

m. Stratify the counties into four strata by population size. What are the sampling fractions for proportional allocation and optimal allocation? Compare the variances of the estimates of the population mean obtained using simple random sampling, proportional allocation, and optimal allocation.

n. How much better than those in part (m) will the estimates of the population mean be if 8, 16, 32, or 64 strata are used instead?

66. A photograph of a large crowd on a beach is taken from a helicopter. The photo is of such high resolution that when sections are magnified, individual people can be identified, but to count the entire crowd in this way would be very time-consuming. Devise a plan to estimate the number of people on the beach by using a sampling procedure.

67. The data set `families` contains information about 43,886 families living in the city of Cyberville. The city has four regions: the Northern region has 10,149 families, the Eastern region has 10,390 families, the Southern region has 13,457 families, and the Western region has 9,890. For each family, the following information is recorded:

1. Family type
1: Husband-wife family
2: Male-head family
3: Female-head family

2. Number of persons in family

3. Number of children in family

4. Family income

5. Region
1: North
2: East
3: South
4: West

6. Education level of head of household
31: Less than 1st grade
32: 1st, 2nd, 3rd, or 4th grade
33: 5th or 6th grade
34: 7th or 8th grade
35: 9th grade
36: 10th grade
37: 11th grade

38: 12th grade, no diploma
39: High school graduate, high school diploma, or equivalent
40: Some college but no degree
41: Associate degree in college (occupation/vocation program)
42: Associate degree in college (academic program)
43: Bachelor's degree (e.g., B.S., B.A., A.B.)
44: Master's degree (e.g., M.S., M.A., M.B.A.)
45: Professional school degree (e.g., M.D., D.D.S., D.V.M., LL.B., J.D.)
46: Doctoral degree (e.g., Ph.D., Ed.D.)

In these exercises, you will try to learn about the families of Cyberville by using sampling.

a. Take a simple random sample of 500 families. Estimate the following population parameters, calculate the estimated standard errors of these estimates, and form 95% confidence intervals:

i. The proportion of female-headed families
ii. The average number of children per family
iii. The proportion of heads of households who did not receive a high school diploma
iv. The average family income

Repeat the preceding parameters for five different simple random samples of size 500 and compare the results.

b. Take 100 samples of size 400.

i. For each sample, find the average family income.
ii. Find the average and standard deviation of these 100 estimates and make a histogram of the estimates.
iii. Superimpose a plot of a normal density with that mean and standard deviation of the histogram and comment on how well it appears to fit.
iv. Plot the empirical cumulative distribution function (see Section 10.2). On this plot, superimpose the normal cumulative distribution function with mean and standard deviation as earlier. Comment on the fit.
v. Another method for examining a normal approximation is via a normal probability plot (Section 9.9). Make such a plot and comment on what it shows about the approximation.
vi. For each of the 100 samples, find a 95% confidence interval for the population average income. How many of those intervals actually contain the population target?
vii. Take 100 samples of size 100. Compare the averages, standard deviations, and histograms to those obtained for a sample of size 400 and explain how the theory of simple random sampling relates to the comparisons.

c. For a simple random sample of 500, compare the incomes of the three family types by comparing histograms and boxplots (see Chapter 10.6).

d. Take simple random samples of size 400 from each of the four regions.

i. Compare the incomes by region by making parallel boxplots.
ii. Does it appear that some regions have larger families than others?
iii. Are there differences in education level among the four regions?

e. Formulate a question of your choice and attempt to answer it with a simple random sample of size 400.

f. Does stratification help in estimating the average family income? From a simple random sample of size 400, estimate the average income and also the standard error of your estimate. Form a 95% confidence interval. Next, allocate the 400 observations proportionally to the four regions and estimate the average income from the stratified sample. Estimate the standard error and form a 95% confidence interval. Compare your results to the results of the simple random sample.