16 Sampling Methods

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- Apply nonprobability sampling methods.
- Decide when to use a stratified sampling method.
- Compute estimates from stratified sample results.
- Decide when to use a cluster sampling method.
- Compute estimates from cluster sampling results.
- Decide when to use a systematic sampling method.
- Compute estimates from systematic sampling results.
- Avoid nonresponse biases in estimates.

16–1 Using Statistics

Throughout this book, we have always assumed that information is obtained through random sampling. The method we have used until now is called simple random sampling. In simple random sampling, we assume that our sample is randomly chosen from the entire population of interest, and that every set of n elements in the population has an equal chance of being selected as our sample.

We assume that a randomization device is always available to the experimenter. We also assume that the entire population of interest is known, in the sense that a random sample can be drawn from the entire population where every element has an equal chance of being included in our sample. Randomly choosing our sample from the entire population in question is our insurance against sampling bias. This was demonstrated in Chapter 5 with the story of the *Literary Digest*.

But do these conditions always hold in real sampling situations? Also, are there any easier ways–more efficient, more economical ways–of drawing random samples? Are there any situations where, instead of randomly drawing every single sample point, we may randomize less frequently and still obtain an adequately random sample for use in statistical inference? Consider the situation where our population is made up of several groups and the elements within each group are similar to one another, but different from the elements in other groups (e.g., sampling an economic variable in a city where some people live in rich neighborhoods and thus form one group, while others live in poorer neighborhoods, forming other groups). Is there a way of using the homogeneity within the groups to make our sampling more efficient? The answers to these questions, as well as many other questions related to sampling methodology, are the subject of this chapter.

In the next section, we will thoroughly explore the idea of random sampling and discuss its practical limitations. We will see that biases may occur if samples are chosen without randomization. We will also discuss criteria for the prevention of such selection biases. In the following sections, we will present more efficient and involved methods of drawing random samples that are appropriate in different situations.

16–2 Nonprobability Sampling and Bias

The advantage of random sampling is that the probabilities that the sample estimator will be within a given number of units from the population parameter it estimates are known. Sampling methods that do not use samples with known probabilities of selection are known as *nonprobability sampling methods*. In such sampling methods, we have no objective way of evaluating how far away from the population parameter our estimate may be. In addition, when we do not select our sample randomly out of the entire population of interest, our sampling results may be biased. That is, the average value of the estimate in repeated sampling is not equal to the parameter of interest. Put simply, our sample may not be a true representative of the population of interest.

EXAMPLE 16–1

A market research firm wants to estimate the proportion of consumers who might be interested in purchasing the Spanish sherry Jerez if this product were available at liquor stores in this country. How should information for this study be obtained?

Solution The population relevant to our case is not a clear-cut one. Before embarking on our study, we need to define our population more precisely. Do we mean all consumers in the United States? Do we mean all families of consumers? Do we mean only people of drinking age? Perhaps we should consider our population to be only those people who, at least occasionally, consume similar drinks. These are important questions to answer before we begin the sampling survey. The population must be defined in accordance with the purpose of the study. In the case of a proposed new product such as Jerez, we are interested in the product's potential market share. We are interested in estimating the proportion of the market for alcoholic beverages that will go to Jerez once it is introduced. Therefore, we define our population as all people who, at least occasionally, consume alcoholic beverages.

Now we need to know how to obtain a random sample from this population. To obtain a random sample out of the whole population of people who drink alcoholic beverages at least occasionally, we must have a *frame*. That is, we need a list of all such people, from which we can randomly choose as many people as we need for our sample. In reality, of course, no such list is available. Therefore, we must obtain our sample in some other way. Market researchers send field workers to places where consumers may be found, usually shopping malls. There shoppers are randomly selected and prescreened to ascertain that they are in the population of interest–in this case, that they are people who at least occasionally consume alcoholic beverages. Then the selected people are given a taste test of the new product and asked to fill out a questionnaire about their response to the product and their future purchase intent. This method of obtaining a random sample works as long as the interviewers do not choose people in a nonrandom fashion, for example, choosing certain types of people because of their appearance. If this should happen, a bias may be introduced if the variable favoring selection is somehow related to interest in the product.

Another point we must consider is the requirement that people selected at the shopping mall constitute a representative sample from the entire population in which we are interested. We must consider the possibility that potential buyers of Jerez may not be found in shopping malls, and if their proportion outside the malls is different from what it is in the malls where surveys take place, a bias will be introduced. We must consider the location of shopping malls and must ascertain that we are not favoring some segments of the population over others. Preferably, several shopping malls, located in different areas, should be chosen.

We should randomize the selection of people or elements in our sample. However, if our sample is not chosen in a purely random way, it may still suffice for our purposes as long as it behaves as a purely random sample and no biases are introduced. In designing the study, we should collect a few random samples at different locations, chosen at different times and handled by different field workers, to minimize the chances of a bias. The results of different samples validate the assumption that we are indeed getting a representative sample.

16–3 Stratified Random Sampling

In some cases, a population may be viewed as comprising different groups where elements in each group are similar to one another in some way. In such cases, we may gain sampling precision (i.e., reduce the variance of our estimators) as well as reduce the costs of the survey by treating the different groups separately. If we consider these groups, or *strata,* as separate subpopulations and draw a separate random sample from each stratum and combine the results, our sampling method is called *stratified random sampling*.

In **stratified random sampling,** we assume that the population of N units may be divided into m groups with N_i units in group i, $i = 1, \ldots, m$. The

m strata are nonoverlapping and together they make up the total population: $N_1 + N_2 + \cdots + N_m = N$.

We define the true *weight* of stratum i as $W_i = N_i/N$. That is, the weight of stratum i is equal to the proportion of the size of stratum i in the whole population. Our total sample, of size n, is divided into subsamples from each of the strata. We sample n_i items in stratum i, and $n_1 + n_2 + \cdots + n_m = n$. We define the sampling fraction in stratum i as $f_i = n_i/N_i$.

The true mean of the entire population is μ, and the true mean in stratum i is μ_i. The variance of stratum i is σ_i^2, and the variance of the entire population is σ^2. The sample mean in stratum i is $\overline{X}_i$, and the combined estimator, the sample mean in stratified random sampling, $\overline{X}_{st}$ is defined as follows:

The **estimator of the population mean** in stratified random sampling is

$$\overline{X}_{st} = \sum_{i=1}^{m} W_i \overline{X}_i \qquad (16\text{–}1)$$

In simple random sampling with no stratification, the stratified estimator in equation 16–1 is, in general, *not equal* to the simple estimator of the population mean. The reason is that the estimator in equation 16–1 uses the true weights of the strata W_i. The simple random sampling estimator of the population mean is $\overline{X} = (\Sigma_{\text{all data}} X)/n = (\Sigma_{i=1}^{\mu} n_i \overline{X}_i)/n$. This is equal to $\overline{X}_{st}$ only if we have $n_i/n = N_i/N$ for each stratum, that is, if the proportion of the sample taken from each stratum is equal to the proportion of each stratum in the entire population. Such a stratification is called stratification with *proportional allocation*.

Following are some important properties of the stratified estimator of the population mean:

1. If the estimator of the mean in each stratum $\overline{X}_i$ is *unbiased,* then the stratified estimator of the mean $\overline{X}_{st}$ is an unbiased estimator of the population mean μ.
2. If the samples in the different strata are drawn *independently* of one another, then the variance of the stratified estimator of the population mean $\overline{X}_{st}$ is given by

$$V(\overline{X}_{st}) = \sum_{i=1}^{m} W_i^2 V(\overline{X}_i) \qquad (16\text{–}2)$$

   where $V(\overline{X}_i)$ is the variance of the sample mean in stratum i.
3. If sampling in all strata is *random,* then the variance of the estimator, given in equation 16–2, is further equal to

$$V(\overline{X}_{st}) = \sum_{i=1}^{m} W_i^2 \left(\frac{\sigma_i^2}{n_i}\right)(1 - f_i) \qquad (16\text{–}3)$$

When the sampling fractions f_i are small and may be ignored, we get

$$V(\overline{X}_{st}) = \sum_{i=1}^{m} W_i^2 \frac{\sigma_i^2}{n_i} \qquad (16\text{–}4)$$

4. If the sample allocation is proportional [$n_i = n(N_i/N)$ for all i], then

$$V(\overline{X}_{st}) = \frac{1-f}{n}\sum_{i=1}^{m} W_i\sigma_i^2 \qquad (16\text{–}5)$$

which reduces to $(1/n)\,\Sigma_{i=1}^{\mu} W_i\sigma_i^2$ when the sampling fraction is small. Note that f is the sampling fraction, that is, the size of the sample divided by the population size.

In addition, if the population variances in all the strata are equal, then

$$V(\overline{X}_{st}) = \frac{\sigma^2}{n} \qquad (16\text{–}6)$$

when the sampling fraction is small.

Practical Applications

In practice, the true population variances in the different strata are usually not known. When the variances are not known, we estimate them from our data. An unbiased estimator of σ_i^2, the population variance in stratum i, is given by

$$S_i^2 = \sum_{\text{data in stratum } i} \frac{(X-\overline{X}_i)^2}{n_i - 1} \qquad (16\text{–}7)$$

The estimator in equation 16–7 is the usual unbiased sample estimator of the population variance in each stratum as a separate population. A particular estimate of the variance in stratum i will be denoted by s_i^2. If sampling in each stratum is random, then an unbiased estimator of the variance of the sample estimator of the population mean is

$$S^2(\overline{X}_{st}) = \sum_{i=1}^{m}\left(\frac{W_i^2 S_i^2}{n_i}\right)(1-f_i) \qquad (16\text{–}8)$$

Any of the preceding formulas apply in the special situations where they can be used with the estimated variances substituted for the population variances.

Confidence Intervals

We now give a confidence interval for the population mean μ obtained from stratified random sampling.

A $(1 - \alpha)$ 100% confidence interval for the population mean μ using stratified sampling is

$$\bar{X}_{st} \pm Z_{\alpha/2} s(\bar{X}_{st}) \qquad (16\text{–}9)$$

where $s(\bar{X}_{st})$ is the square root of the estimate of the variance of $\bar{X}_{st}$ given in equation 16–8.

When the sample sizes in some of the strata are small and the population variances are unknown, but the populations are at least approximately normal, we use the t distribution instead of Z. We denote the degrees of freedom of the t distribution by df. The exact value of df is difficult to determine, but it lies somewhere between the smallest $n_i - 1$ (the degrees of freedom associated with the sample from stratum i) and the sum of the degrees of freedom associated with the samples from all strata $\Sigma_{i=1}^{m}(n_i - 1)$. An approximation for the effective number of degrees of freedom is given by

$$\text{Effective df} = \frac{\left[\sum_{i=1}^{m} N_i(N_i - n_i)s_i^2/n_i\right]^2}{\sum_{i=1}^{m}[N_i(N_i - n_i)/n_i]^2 s_i^4/(n_i - 1)} \qquad (16\text{–}10)$$

We demonstrate the application of the theory of stratified random sampling presented so far by the following example.

EXAMPLE 16–2

Once a year, *Fortune* magazine publishes the Fortune Service 500, a list of the largest service companies in the United States. The 500 firms belong to six major industry groups. The industry groups and the number of firms in each group are listed in Table 16–1.

The 500 firms are considered a complete population: the population of the top 500 service companies in the United States. An economist who is interested in this population wants to estimate the mean net income of all firms in the index. However, obtaining the data for all 500 firms in the index is difficult, time-consuming, or costly. Therefore, the economist wants to gather a random sample of the firms, compute a quick average of the net income for the firms in the sample, and use it to estimate the mean net income for the entire population of 500 firms.

TABLE 16–1 The Fortune Service 500

Group	Number of Firms
1. Diversified service companies	100
2. Commercial banking companies	100
3. Financial service companies (including savings and insurance)	150
4. Retailing companies	50
5. Transportation companies	50
6. Utilities	50
	500

Solution The economist believes that firms in the same industry group share common characteristics related to net income. Therefore, the six groups are treated as different strata, and a random sample is drawn from each stratum. The weights of each of the strata are known exactly as computed from the strata sizes in Table 16–1. Using the definition of the population weights $W_i = N_i/N$, we get the following weights:

$$\begin{aligned} W_1 &= N_1/N = 100/500 = 0.2 \\ W_2 &= N_2/N = 100/500 = 0.2 \\ W_3 &= N_3/N = 150/500 = 0.3 \\ W_4 &= N_4/N = 50/500 = 0.1 \\ W_5 &= N_5/N = 50/500 = 0.1 \\ W_6 &= N_6/N = 50/500 = 0.1 \end{aligned}$$

The economist decides to select a random sample of 100 of the 500 firms listed in *Fortune*. The economist chooses to use a proportional allocation of the total sample to the six strata (another method of allocation will be presented shortly). With proportional allocation, the total sample of 100 must be allocated to the different strata in proportion to the computed strata weights. Thus, for each i, $i = 1, \ldots, 6$, we compute n_i as $n_i = nW_i$. This gives the following sample sizes:

$$n_1 = 20 \quad n_2 = 20 \quad n_3 = 30 \quad n_4 = 10 \quad n_5 = 10 \quad n_6 = 10$$

We will assume that the net income values in the different strata are approximately normally distributed and that the estimated strata variances (to be estimated from the data) are the true strata variances σ_i^2, so that the normal distribution may be used.

The economist draws the random samples and computes the sample means and variances. The results, in millions of dollars (for the means) and in millions of dollars squared (for the variances), are given in Table 16–2, along with the sample sizes in the different strata and the strata weights. From the table, and with the aid of equation 16–1 for the mean and equation 16–5 for the variance of the sample mean (with the estimated sample variances substituted for the population variances in the different strata), we will now compute the stratified sample mean and the estimated variance of the stratified sample mean.

$$\begin{aligned} \bar{x}_{st} = \sum_{i=1}^{6} W_i \bar{x}_i &= (0.2)(52.7) + (0.2)(112.6) + (0.3)(85.6) + (0.1)(12.6) \\ &\quad + (0.1)(8.9) + (0.1)(52.3) \\ &= \$66.12 \text{ million} \end{aligned}$$

and

$$\begin{aligned} s(\bar{X}_{st}) &= \sqrt{\frac{1-f}{n}\sum_{i=1}^{6} W_i s_i^2} \\ &= \sqrt{\frac{0.8}{100}[(0.2)(97{,}650) + (0.2)(64{,}300) + (0.3)(76{,}990) + (0.1)(18{,}320)} \\ &\quad \overline{+ (0.1)(9{,}037) + (0.1)(83{,}500)]} \\ &= 23.08 \end{aligned}$$

TABLE 16–2 Sampling Results for Example 16–2

Stratum	Mean	Variance	n_i	W_i
1	52.7	97,650	20	0.2
2	112.6	64,300	20	0.2
3	85.6	76,990	30	0.3
4	12.6	18,320	10	0.1
5	8.9	9,037	10	0.1
6	52.3	83,500	10	0.1

(Our sampling fraction is $f = 100/500 = 0.2$.) Our unbiased point estimate of the average net income of all firms in the Fortune Service 500 is \$66.12 million.

Using equation 16–9, we now compute a 95% confidence interval for μ, the mean net income of all firms in the index. We have the following:

$$\bar{x}_{st} \pm z_{\alpha/2} s(\bar{X}_{st}) = 66.12 \pm (1.96)(23.08) = [20.88, 111.36]$$

Thus, the economist may be 95% confident that the average net income for all firms in the Fortune Service 500 is anywhere from 20.88 to 111.36 million dollars. Incidentally, the true population mean net income for all 500 firms in the index is μ = \$61.496 million.

The Template

Figure 16–1 shows the template that can be used to estimate population means by stratified sampling in the example.

Stratified Sampling for the Population Proportion

The theory of stratified sampling extends in a natural way to sampling for the population proportion p. Let the sample proportion in stratum i be $\hat{P}_i = X_i/n_i$, where X_i is the number of successes in a sample of size n_i. Then the stratified estimator of the population proportion p is the following.

FIGURE 16–1 The Template for Estimating Means by Stratified Sampling [Stratified Sampling.xls; Sheet: Mean]

Stratified Sampling for Estimating μ

Fortune Service 500

Stratum	*N*	*n*	*x*-bar	s^2
1	100	20	52.7	97650
2	100	20	112.6	64300
3	150	30	85.6	76990
4	50	10	12.6	18320
5	50	10	8.9	9037
6	50	10	52.3	83500
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
Total	500	100		

X-bar	V(*X*-bar)	S(*X*-bar)	*df*
66.12	532.5816	23.07773	80

1 − α	(1 − α) CI for *X*-bar				
95%	66.12	+ or -	45.92614	or	20.19386 to 112.0461

The stratified estimator of the population proportion p is

$$\hat{P}_{st} = \sum_{i=1}^{m} W_i \hat{P}_i \qquad (16\text{–}11)$$

where the weights W_i are defined as in the case of sampling for the population mean: $W_i = N_i/N$.

The following is an approximate expression for the variance of the estimator of the population proportion $\hat{P}_{st}$, for use with large samples.

The approximate variance of $\hat{P}_{st}$ is

$$V(\hat{P}_{st}) = \sum_{i=1}^{m} W_i^2 \frac{\hat{P}_i \hat{Q}_i}{n_i} \qquad (16\text{–}12)$$

where $\hat{Q}_i = 1 - \hat{P}_i$.

When finite-population correction factors f_i must be considered, the following expression is appropriate for the variance of $\hat{P}_{st}$:

$$V(\hat{P}_{st}) = \frac{1}{N^2} \sum_{i=1}^{m} N_i^2 (N_i - n_i) \frac{\hat{P}_i \hat{Q}_i}{(N_i - 1) n_i} \qquad (16\text{–}13)$$

When proportional allocation is used, an approximate expression is

$$V(\hat{P}_{st}) = \frac{1 - f}{n} \sum_{i=1}^{m} W_i \hat{P}_i \hat{Q}_i \qquad (16\text{–}14)$$

Let us now return to Example 16–1, sampling for the proportion of people who might be interested in purchasing the Spanish sherry Jerez. Suppose that the market researchers believe that preferences for imported wines differ between consumers in metropolitan areas and those in other areas. The area of interest for the survey covers a few states in the Northeast, where it is known that 65% of the people live in metropolitan areas and 35% live in nonmetropolitan areas. A sample of 130 people randomly chosen at shopping malls in metropolitan areas shows that 28 are interested in Jerez, while a random sample of 70 people selected at malls outside the metropolitan areas shows that 18 are interested in the sherry.

Let us use these results in constructing a 90% confidence interval for the proportion of people in the entire population who are interested in the product. From equation 16–11, using two strata with weights 0.65 and 0.35, we get

$$\hat{p}_{st} = \sum_{i=1}^{2} W_i \hat{p}_i = (0.65)\frac{28}{130} + (0.35)\frac{18}{70} = 0.23$$

Our allocation is proportional because $n_1 = 130$, $n_2 = 70$, and $n = 130 + 70 = 200$, so that $n_1/n = 0.65 = W_1$ and $n_2/n = 0.35 = W_2$. In addition, the sample sizes of 130 and 70 represent tiny fractions of the two strata; hence, no finite-population correction factor is required. The equation for the estimated variance of the sample estimator of the proportion is therefore equation 16–14 without the finite-population correction:

$$V(\hat{P}_{st}) = \frac{1}{n}\sum_{i=1}^{2} W_i \hat{p}_i \hat{q}_i = \frac{1}{200}[(0.65)(0.215)(0.785) + (0.35)(0.257)(0.743)]$$
$$= 0.0008825$$

The standard error of $\hat{P}_{st}$ is therefore $\sqrt{V(\hat{P}_{st})} = \sqrt{0.0008825} = 0.0297$. Thus, our 90% confidence interval for the population proportion of people interested in Jerez is

$$\hat{p}_{st} \pm z_{\alpha/2}\, s(\hat{P}_{st}) = 0.23 \pm (1.645)(0.0297) = [0.181, 0.279]$$

The stratified point estimate of the percentage of people in the proposed market area for Jerez who may be interested in the product, if it is introduced, is 23%. A 90% confidence interval for the population percentage is 18.1% to 27.9%.

The Template

Figure 16–2 shows the template that can be used to estimate population proportions by stratified sampling. The data in the figure correspond to the Jerez example.

What Do We Do When the Population Strata Weights Are Unknown?

Here and in the following subsections, we will explore some optional, advanced aspects of stratified sampling. When the true strata weights $W_i = N_i/N$ are unknown–that is, when we do not know what percentage of the whole population belongs to each

FIGURE 16–2 The Template for Estimating Proportions by Stratified Sampling [Stratified Sampling.xls; Sheet: Proportion]

Stratified Sampling for Estimating Proportion — Jerez

Stratum	*N*	*n*	*x*	*p*-hat
1	650000	130	28	0.21538
2	350000	70	18	0.25714
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
Total	1000000	200		

P-hat	V(*P*-hat)	S(*P*-hat)
0.2300	0.000883	0.0297

$1-\alpha$	$(1-\alpha)$ CI for *P*				
90%	0.2300	+ or -	0.0489	or	0.1811 to 0.2789

stratum—we may still use stratified random sampling. In such cases, we use estimates of the true weights, denoted by w_i. The consequence of using estimated weights instead of the true weights is the introduction of a bias into our sampling results. The interesting thing about this kind of bias is that it is not eliminated as the sample size increases. When errors in the stratum weights exist, our results are always biased; the greater the errors, the greater the bias. These errors also cause the standard error of the sample mean $s(\bar{X}_{st})$ to underestimate the true standard deviation. Consequently, confidence intervals for the population parameter of interest tend to be narrower than they should be.

How Many Strata Should We Use?

The number of strata to be used is an important question to consider when you are designing any survey that uses stratified sampling. In many cases, there is a natural breakdown of the population into a given number of strata. In other cases, there may be no clear, unique way of separating the population into groups. For example, if age is to be used as a stratifying variable, there are many ways of breaking the variable and forming strata. The two guidance rules for constructing strata are presented below.

Rules for Constructing Strata

1. The number of strata should preferably be less than or equal to 6.
2. Choose the strata so that Cum $\sqrt{f(x)}$ is approximately constant for all strata [where Cum $\sqrt{f(x)}$ is the cumulative square root of the frequency of X, the variable of interest].

The first rule is clear. The second rule says that in the absence of other guidelines for breaking down a population into strata, we partition the variable used for stratification into categories so that the cumulative square root of the frequency function of the variable is approximately equal for all strata. We illustrate this rule in the hypothetical case of Table 16–3. As can be seen in this simplified example, the combined age groups 20–30, 31–35, and 36–45 all have a sum of $\sqrt{f}$ equal to 5; hence, these groups make good strata with respect to age as a stratifying variable according to rule 2.

Postsampling Stratification

At times, we conduct a survey using simple random sampling with no stratification, and after obtaining our results, we note that the data may be broken into categories of similar elements. Can we now use the techniques of stratified random sampling and enjoy its benefits in terms of reduced variances of the estimators? Surprisingly, the answer is yes. In fact, if the subsamples in each of our strata contain at least 20 elements, and if our estimated weights of the different strata w_i (computed from the data as n_i/n, or from more accurate information) are close to the true population strata weights W_i, then our stratified estimator will be almost as good as that of stratified random sampling with proportional allocation. This procedure is called *poststratification*.

TABLE 16–3 Constructing Strata by Age

Age	Frequency f	$\sqrt{f}$	Cum $\sqrt{f}$
20–25	1	1	
26–30	16	4	5
31–35	25	5	5
36–40	4	2	
41–45	9	3	5

We close this section with a discussion of an alternative to proportional allocation of the sample in stratified random sampling, called *optimum allocation.*

Optimum Allocation

With optimum allocation, we select the sample sizes to be allocated to each of the strata so as to minimize one of two criteria. Either we minimize the cost of the survey for a given value of the variance of our estimator, or we minimize the variance of our estimator for a given cost of taking the survey.

We assume a cost function of the form

$$C = C_0 + \sum_{i=1}^{m} C_i n_i \qquad (16\text{–}15)$$

where C is the total cost of the survey, C_0 is the fixed cost of setting up the survey, and C_i is the cost per item sampled in stratum i. Clearly, the total cost of the survey is the sum of the fixed cost and the costs of sampling in all the strata (where the cost of sampling in stratum i is equal to the sample size n_i times the cost per item sampled C_i).

Under the assumption of a cost function given in equation 16–15, the optimum allocation that will minimize our total cost for a fixed variance of the estimator, or minimize the variance of the estimator for a fixed total cost, is as follows.

Optimum allocation:

$$\frac{n_i}{n} = \frac{W_i \sigma_i / \sqrt{C_i}}{\sum_{i=1}^{m} W_i \sigma_i / \sqrt{C_i}} \qquad (16\text{–}16)$$

Equation 16–16 has an intuitive appeal. It says that for a given stratum, we should take a larger sample if the stratum is *more variable internally* (greater σ_i), if the relative *size of the stratum is larger* (greater W_i), or if *sampling in the stratum is cheaper* (smaller C_i).

If the cost per unit sampled is the same in all the strata (i.e., if $C_i = c$ for all i), then the optimum allocation for a fixed total cost is the same as the optimum allocation for fixed sample size, and we have what is called the *Neyman allocation* (after J. Neyman, although this allocation was actually discovered earlier by A. A. Tschuprow in 1923).

The Neyman allocation:

$$\frac{n_i}{n} = \frac{W_i \sigma_i}{\sum_{i=1}^{m} W_i \sigma_i} \qquad (16\text{–}17)$$

Suppose that we want to allocate a total sample of size 1,000 to three strata, where stratum 1 has weight 0.4, standard deviation 1, and cost per sampled item of 4 cents; stratum 2 has weight 0.5, standard deviation 2, and cost per item of 9 cents;

and stratum 3 has weight 0.1, standard deviation 3, and cost per item of 16 cents. How should we allocate this sample if optimum allocation is to be used? We have

$$\sum_{i=1}^{3} \frac{W_i \sigma_i}{\sqrt{C_i}} = \frac{(0.4)(1)}{\sqrt{4}} + \frac{(0.5)(2)}{\sqrt{9}} + \frac{(0.1)(3)}{\sqrt{16}} = 0.608$$

From equation 16–16, we get

$$\frac{n_1}{n} = \frac{W_1\sigma_1/\sqrt{C_1}}{\sum_{i=1}^{3}(W_i\sigma_i/\sqrt{C_i})} = \frac{(0.4)(1)/\sqrt{4}}{0.608} = 0.329$$

$$\frac{n_2}{n} = \frac{W_2\sigma_2/\sqrt{C_2}}{\sum_{i=1}^{3}(W_i\sigma_i/\sqrt{C_i})} = \frac{(0.5)(2)/\sqrt{9}}{0.608} = 0.548$$

$$\frac{n_3}{n} = \frac{W_3\sigma_3/\sqrt{C_3}}{\sum_{i=1}^{3}(W_i\sigma_i/\sqrt{C_i})} = \frac{(0.1)(3)/\sqrt{16}}{0.608} = 0.123$$

The optimum allocation in this case is 329 items from stratum 1; 548 items from stratum 2; and 123 items from stratum 3 (making a total of 1,000 sample items, as specified).

Let us now compare this allocation with proportional allocation. With a sample of size 1,000 and a proportional allocation, we would allocate our sample only by the stratum weights, which are 0.4, 0.5, and 0.1, respectively. Therefore, our allocation will be 400 from stratum 1; 500 from stratum 2; and 100 from stratum 3. The optimum allocation is different, as it incorporates the cost and variance considerations. Here, the difference between the two sets of sample sizes is not large.

Suppose, in this example, that the costs of sampling from the three strata are the same. In this case, we can use the Neyman allocation and get, from equation 16–17,

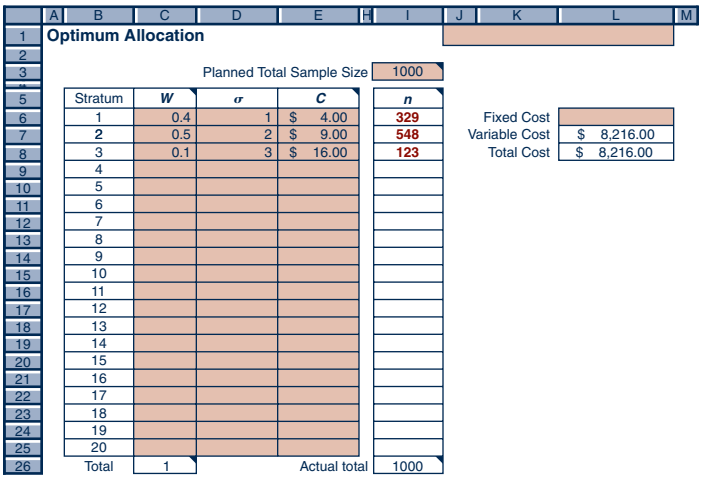
$$\frac{n_1}{n} = \frac{W_1\sigma_1}{\sum_{i=1}^{3} W_i\sigma_i} = \frac{(0.4)(1)}{1.7} = 0.235$$

$$\frac{n_2}{n} = \frac{W_2\sigma_2}{\sum_{i=1}^{3} W_i\sigma_i} = \frac{(0.5)(2)}{1.7} = 0.588$$

$$\frac{n_3}{n} = \frac{W_3\sigma_3}{\sum_{i=1}^{3} W_i\sigma_i} = \frac{(0.1)(3)}{1.7} = 0.176$$

Thus, the Neyman allocation gives a sample of size 235 to stratum 1; 588 to stratum 2; and 176 to stratum 3. Note that these subsamples add only to 999, due to rounding error. The last sample point may be allocated to any of the strata.

FIGURE 16–3 The Template for Optimal Allocation for Stratified Sampling [Stratified Sampling.xls; Sheet: Allocation]



Optimum Allocation

Planned Total Sample Size: 1000

Stratum	W	σ	C	n
1	0.4	1	$ 4.00	329
2	0.5	2	$ 9.00	548
3	0.1	3	$ 16.00	123
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
Total	1		Actual total	1000

Fixed Cost	
Variable Cost	$ 8,216.00
Total Cost	$ 8,216.00

In general, stratified random sampling gives more precise results than those obtained from simple random sampling: The standard errors of our estimators from stratified random sampling are usually smaller than those of simple random sampling. Furthermore, in stratified random sampling, an optimum allocation will produce more precise results than a proportional allocation if some strata are more expensive to sample than others or if the variances within strata are different from one another.

The Template

Figure 16–3 shows the template that can be used to estimate population proportions by stratified sampling. The data in the figure correspond to the example we have been discussing.

The same template can be used for Neyman allocation. In column E, enter the same cost *C*, say $1, for all the strata.

PROBLEMS

16–1. A securities analyst wants to estimate the average percentage of institutional holding of all publicly traded stocks in the United States. The analyst believes that stocks traded on the three major exchanges have different characteristics and therefore decides to use stratified random sampling. The three strata are the New York Stock Exchange (NYSE), the American Exchange (AMEX), and the Over the Counter (OTC) exchange. The weights of the three strata, as measured by the number of stocks listed in each exchange, divided by the total number of stocks, are NYSE, 0.44; AMEX, 0.15; OTC, 0.41. A total random sample of 200 stocks is selected, with proportional allocation. The average percentage of institutional holdings of the subsample selected from the issues of the NYSE is 46%, and the standard deviation is 8%. The corresponding results for the AMEX are 9% average institutional holdings and a standard deviation of 4%, and the corresponding results for the OTC stocks are 29% average institutional holdings and a standard deviation of 16%.

a. Give a stratified estimate of the mean percentage of institutional holdings per stock.

b. Give the standard error of the estimate in *a*.

c. Give a 95% confidence interval for the mean percentage of institutional holdings.

d. Explain the advantages of using stratified random sampling in this case. Compare with simple random sampling.

16–2. A company has 2,100 employees belonging to the following groups: production, 1,200; marketing, 600; management, 100; other, 200. The company president wants to obtain an estimate of the views of all employees about a certain impending executive decision. The president knows that the management employees' views are most variable, along with employees in the "other" category, while the marketing and production people have rather uniform views within their groups. The production people are the most costly to sample, because of the time required to find them at their different jobs, and the management people are easiest to sample.

a. Suppose that a total sample of 100 employees is required. What are the sample sizes in the different strata under proportional allocation?

b. Discuss how you would design an optimum allocation in this case.

16–3. Last year, consumers increasingly bought fleece (industry jargon for hot-selling jogging suits, which now rival jeans as the uniform for casual attire). A New York designer of jogging suits is interested in the new trend and wants to estimate the amount spent per person on jogging suits during the year. The designer knows that people who belong to health and fitness clubs will have different buying behavior than people who do not. Furthermore, the designer finds that, within the proposed study area, 18% of the population are members of health and fitness clubs. A random sample of 300 people is selected, and the sample is proportionally allocated to the two strata: members of health clubs and nonmembers of health clubs. It is found that among members, the average amount spent is \$152.43 and the standard deviation is \$25.77, while among the nonmembers, the average amount spent is \$15.33 and the standard deviation is \$5.11.

a. What is the stratified estimate of the mean?

b. What is the standard error of the estimator?

c. Give a 90% confidence interval for the population mean μ.

d. Discuss one possible problem with the data. (*Hint:* Can the data be considered normally distributed? Why?)

16–4. A financial analyst is interested in estimating the average amount of a foreign loan by U.S. banks. The analyst believes that the amount of a loan may be different depending on the bank, or, more precisely, on the extent of the bank's involvement in foreign loans. The analyst obtains the following data on the percentage of profits of U.S. banks from loans to Mexico and proposes to use these data in the construction of stratum weights. The strata are the different banks: First Chicago, 33%; Manufacturers Hanover, 27%; Bankers Trust, 21%; Chemical Bank, 19%; Wells Fargo Bank, 19%; Citicorp, 16%; Mellon Bank, 16%; Chase Manhattan, 15%; Morgan Guarantee Trust, 9%.

a. Construct the stratum weights for proportional allocation.

b. Discuss two possible problems with this study.

16–4 Cluster Sampling

Let us consider the case where we have no frame (i.e., no list of all the elements in the population) and the elements are *clustered* in larger units. Each unit, or cluster, contains several elements of the population. In this case, we may choose to use the method of **cluster sampling.** This may also be the case when the population is large and spread over a geographic area in which smaller subregions are easily sampled and where a simple random sample or a stratified random sample may not be carried out as easily.

Suppose that the population is composed of M clusters and there is a list of all M clusters from which a random sample of m clusters is selected. Two possibilities arise. First, we may sample *every element* in every one of the m selected clusters. In this case, our sampling method is called *single-stage cluster sampling*. Second, we may select a random sample of m clusters and then select a random sample of n elements from each of the selected clusters. In this case, our sampling method is called *two-stage cluster sampling*.

The Relation with Stratified Sampling

In stratified sampling, we sample elements from every one of our strata, and this assures us of full representation of all segments of the population in the sample. In cluster sampling, we sample only some of the clusters, and although elements within any cluster may tend to be homogeneous, as is the case with strata, not all the clusters are represented in the sample; this leads to lowered precision of the cluster sampling method. In stratified random sampling, we use the fact that the population may be broken into subgroups. This usually leads to a smaller variance of our estimators. In cluster sampling, however, the method is used mainly because of ease of implementation or reduction in sampling costs, and the estimates do not usually lead to more precise results.

Single-Stage Cluster Sampling for the Population Mean

Let $n_1, n_2, \ldots, n_m$ be the number of elements in each of the m sampled clusters. Let $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_m$ be the means of the sampled clusters. The cluster sampling unbiased estimator of the population mean μ is given as follows.

The cluster sampling estimator of μ is

$$\bar{X}_{\text{cl}} = \frac{\sum_{i=1}^{m} n_i \bar{X}_i}{\sum_{i=1}^{m} n_i} \tag{16–18}$$

An estimator of the variance of the estimator of μ in equation 16–18 is

$$s^2(\bar{X}_{\text{cl}}) = \frac{M - m}{Mm\bar{n}^2} \frac{\sum_{i=1}^{m} n_i^2 (\bar{X}_i - \bar{X}_{\text{cl}})^2}{m - 1} \tag{16–19}$$

where $\bar{n} = (\Sigma_{i=1}^{m} n_i)/m$ is the average number of units in the sampled clusters.

Single-Stage Cluster Sampling for the Population Proportion

The cluster sampling estimator of the population proportion p is

$$\hat{P}_{\text{cl}} = \frac{\sum_{i=1}^{m} n_i \hat{P}_i}{\sum_{i=1}^{m} n_i} \tag{16–20}$$

where the $\hat{P}_i$ are the proportions of interest within the sampled clusters.

The estimated variance of the estimator in equation 16–20 is given by

$$s^2(\hat{P}_{\text{cl}}) = \frac{M - m}{Mm\bar{n}^2} \frac{\sum_{i=1}^{m} n_i^2(\hat{P}_i - \hat{P}_{\text{cl}})^2}{m - 1} \tag{16–21}$$

We now demonstrate the use of cluster sampling for the population mean with the following example.

EXAMPLE 16–3

J. B. Hunt Transport Company is especially interested in lowering fuel costs in order to survive in the tough world of deregulated trucking. Recently, the company introduced new measures to reduce fuel costs for all its trucks. Suppose that company trucks are based in 110 centers throughout the country and that the company's management wants to estimate the average amount of fuel saved per truck for the week following the institution of the new measures. For reasons of lower cost and administrative ease, management decides to use single-stage cluster sampling, select a random sample of 20 trucking centers, and measure the weekly fuel saving for each of the trucks in the selected centers (each center is a cluster). The average fuel savings per truck, in gallons, for each of the 20 selected centers are as follows (the number of trucks in each center is given in parentheses): 21 (8), 22 (8), 11 (9), 34 (10), 28 (7), 25 (8), 18 (10), 24 (12), 19 (11), 20 (6), 30 (8), 26 (9), 12 (9), 17 (8), 13 (10), 29 (8), 24 (8), 26 (10), 18 (10), 22 (11). From these data, compute an estimate of the average amount of fuel saved per truck for all Hunt's trucks over the week in question. Also give a 95% confidence interval for this parameter.

Solution From equation 16–18, we get

$$\begin{aligned}\bar{x}_{\text{cl}} = \frac{\sum_{i=1}^{m} n_i\bar{x}_i}{\sum_{i=1}^{m} n_i} = {} & [21(8) + 22(8) + 11(9) + 34(10) + 28(7) + 25(8) + 18(10) \\ & + 24(12) + 19(11) + 20(6) + 30(8) + 26(9) + 12(9) + 17(8) \\ & + 13(10) + 29(8) + 24(8) + 26(10) + 18(10) + 22(11)]/(8 \\ & + 8 + 9 + 10 + 7 + 8 + 10 + 12 + 11 + 6 + 8 + 9 + 9 \\ & + 8 + 10 + 8 + 8 + 10 + 10 + 11) \\ = {} & 21.83\end{aligned}$$

From equation 16–19, we find that the estimated variance of our sample estimator of the mean is

$$\begin{aligned}s^2(\bar{X}_{\text{cl}}) &= \frac{M - m}{Mm\bar{n}^2} \frac{\sum_{i=1}^{20} n_i^2(\bar{x}_i - \bar{x}_{\text{cl}})^2}{m - 1} \\ &= \frac{110 - 20}{(110)(20)(9)^2}[8^2(21 - 21.83)^2 + 8^2(22 - 21.83)^2 \\ &\qquad + 9^2(11 - 21.83)^2 + \cdots + 11^2(22 - 21.83)^2]/19 \\ &= 1.587\end{aligned}$$

Using the preceding information, we construct a 95% confidence interval for μ as follows:

$$\bar{x}_{\text{cl}} \pm 1.96s(\bar{X}_{\text{cl}}) = 21.83 \pm 1.96\sqrt{1.587} = [19.36, 24.30]$$

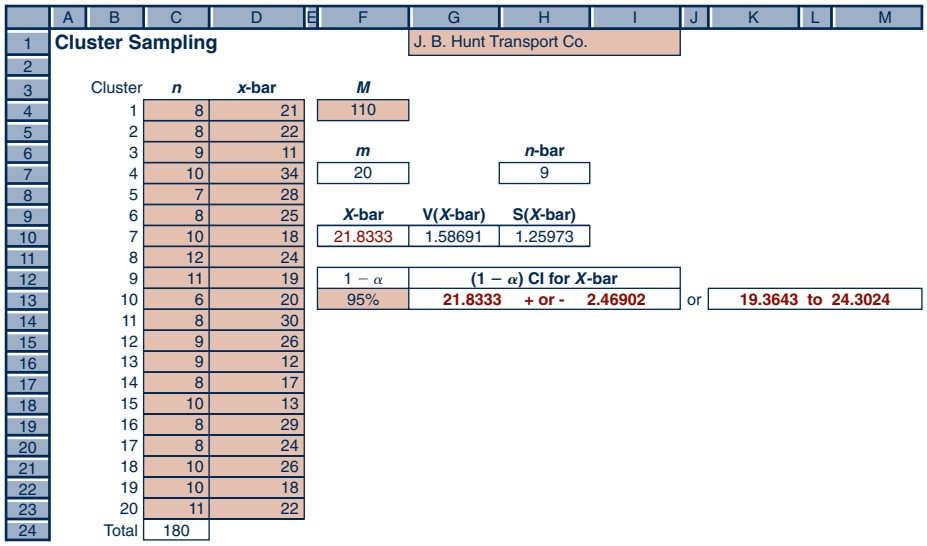
Thus, based on the sampling results, Hunt's management may be 95% confident that average fuel savings per truck for all trucks over the week in question is anywhere from 19.36 to 24.30 gallons.

The Templates

The template for estimating a population mean by cluster sampling is shown in Figure 16–4. The data in the figure correspond to Example 16–3.

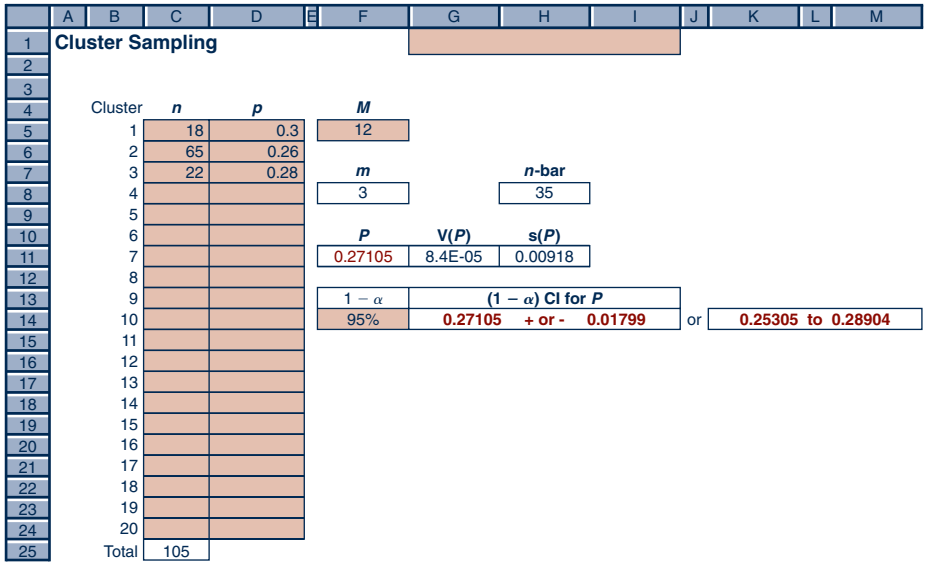
The template for estimating a population proportion by cluster sampling is shown in Figure 16–5.

FIGURE 16–4 The Template for Estimating Means by Cluster Sampling [Cluster Sampling.xls; Sheet: Mean]



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Cluster Sampling**						J. B. Hunt Transport Co.						
2													
3		Cluster	***n***	***x*-bar**		***M***							
4		1	8	21		110							
5		2	8	22									
6		3	9	11		***m***		***n*-bar**					
7		4	10	34		20		9					
8		5	7	28									
9		6	8	25		***X*-bar**	**V(*X*-bar)**	**S(*X*-bar)**					
10		7	10	18		21.8333	1.58691	1.25973					
11		8	12	24									
12		9	11	19		$1-\alpha$	**(1 − α) CI for *X*-bar**						
13		10	6	20		95%	**21.8333 + or - 2.46902**			or	**19.3643 to 24.3024**		
14		11	8	30									
15		12	9	26									
16		13	9	12									
17		14	8	17									
18		15	10	13									
19		16	8	29									
20		17	8	24									
21		18	10	26									
22		19	10	18									
23		20	11	22									
24		Total	180										

FIGURE 16–5 The Template for Estimating Proportions by Cluster Sampling [Cluster Sampling.xls; Sheet: Proportion]



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Cluster Sampling**												
2													
3													
4		Cluster	***n***	***p***		***M***							
5		1	18	0.3		12							
6		2	65	0.26									
7		3	22	0.28		***m***		***n*-bar**					
8		4				3		35					
9		5											
10		6				***P***	**V(*P*)**	**s(*P*)**					
11		7				0.27105	8.4E-05	0.00918					
12		8											
13		9				$1-\alpha$	**(1 − α) CI for *P***						
14		10				95%	**0.27105 + or - 0.01799**			or	**0.25305 to 0.28904**		
15		11											
16		12											
17		13											
18		14											
19		15											
20		16											
21		17											
22		18											
23		19											
24		20											
25		Total	105										

Two-Stage Cluster Sampling

When clusters are very large or when elements within each cluster tend to be similar, we may gain little information by sampling every element within the selected clusters. In such cases, selecting more clusters and sampling only some of the elements within the chosen clusters may be more economical. The formulas for the estimators and their variances in the case of two-stage cluster sampling are more complicated and may be found in advanced books on sampling methodology.

PROBLEMS

16–5. There are 602 aerobics and fitness centers in Japan (up from 170 five years ago). Adidas, the European maker of sports shoes and apparel, is very interested in this fast-growing potential market for its products. As part of a marketing survey, Adidas wants to estimate the average income of all members of Japanese fitness centers. (Members of one such club pay $62 to join and another $62 per month. Adidas believes that the average income of all fitness club members in Japan may be higher than that of the general population, for which census data exist.) Since travel and administrative costs for conducting a simple random sample of all members of fitness clubs throughout Japan would be prohibitive, Adidas decided to conduct a cluster sampling survey. Five clubs were chosen at random out of the entire collection of 602 clubs, and all members of the five clubs were interviewed. The following are the average incomes (in U.S. dollars) for the members of each of the five clubs (the number of members in each club is given in parentheses): $37,237 (560), $41,338 (435), $28,800 (890), $35,498 (711), $47,446 (230). Give the cluster sampling estimate of the population mean income for all fitness club members in Japan. Also give a 90% confidence interval for the population mean. Are there any limitations to the methodology in this case?

16–6. Israel's kibbutzim are by now well diversified beyond their agrarian roots, producing everything from lollipops to plastic pipe. These 282 widely scattered communes of several hundred members maintain hundreds of factories and other production facilities. An economist wants to estimate the average annual revenues of all kibbutz production facilities. Since each kibbutz has several production units, and since travel and other costs are high, the economist wants to consider a sample of 15 randomly chosen kibbutzim and find the annual revenues of all production units in the selected kibbutzim. From these data, the economist hopes to estimate the average annual revenue per production unit in all 282 kibbutzim. The sample results are as follows:

Kibbutz	Number of Production Units	Total Kibbutz Annual Revenues (in millions of dollars)
1	4	4.5
2	2	2.8
3	6	8.9
4	2	1.2
5	5	7.0
6	3	2.2
7	2	2.3
8	1	0.8
9	8	12.5
10	4	6.2
11	3	5.5
12	3	6.2
13	2	3.8
14	5	9.0
15	2	1.4

From these data, compute the cluster sampling estimate of the mean annual revenue of all kibbutzim production units, and give a 95% confidence interval for the mean.

16–7. Under what conditions would you use cluster sampling? Explain the differences among cluster sampling, simple random sampling, and stratified random sampling. Under what conditions would you use two-stage cluster sampling? Explain the difference between single-stage and two-stage cluster sampling. What are the limitations of cluster sampling?

16–8. Recently a survey was conducted to assess the quality of investment brokers. A random sample of 6 brokerage houses was selected from a total of 27 brokerage houses. Each of the brokers in the selected brokerage houses was evaluated by an independent panel of industry experts as "highly qualified" (HQ) or was given an evaluation below this rating. The designers of the survey wanted to estimate the proportion of all brokers in the entire industry who would be considered highly qualified. The survey results are in the following table.

Brokerage House	Total Number of Brokers	Number of HQ Brokers
1	120	80
2	150	75
3	200	100
4	100	65
5	88	45
6	260	200

Use the cluster sampling estimator of the population proportion to estimate the proportion of all highly qualified brokers in the investment industry. Also give a 99% confidence interval for the population proportion you estimated.

16–9. Forty-two cruise ships come to Alaska's Glacier Bay every year. The state tourist board wants to estimate the average cruise passenger's satisfaction from this experience, rated on a scale of 0 to 100. Since the ships' arrivals are evenly spread throughout the season, simple random sampling is costly and time-consuming. Therefore, the agency decides to send its volunteers to board the first five ships of the season, consider them as clusters, and randomly choose 50 passengers in each ship for interviewing.

a. Is the method employed single-stage cluster sampling? Explain.

b. Is the method employed two-stage cluster sampling? Explain.

c. Suppose that each of the ships has exactly 50 passengers. Is the proposed method single-stage cluster sampling?

d. The 42 ships belong to 12 cruise ship companies. Each company has its own characteristics in terms of price, luxury, services, and type of passengers. Suggest an alternative sampling method, and explain its benefits.

16–5 Systematic Sampling

Sometimes a population is arranged in some order: files in a cabinet, crops in a field, goods in a warehouse, etc. In such cases drawing our random sample in a *systematic* way may be easier than generating a simple random sample that would entail looking for particular items within the population. To select a **systematic sample** of n elements from a population of N elements, we divide the N elements in the population into n groups of k elements and then use the following rule:

We randomly select the first element out of the first k elements in the population, and then we select every kth unit afterward until we have a sample of n elements.

For example, suppose $k = 20$, and we need a sample of $n = 30$ items. We randomly select the first item from the integers 1 to 20. If the random number selected is 11, then our systematic sample will contain the elements 11, $11 + 20 = 31$, $31 + 20 = 51$, . . . , and so on until we have 30 elements in our sample.

A variant of this rule, which solves the problems that may be encountered when k is not an integer multiple of the sample size n (their product being N), is to let k be the nearest integer to N/n. We now regard the N elements as being arranged in a circle (with the last element preceding the first element). We randomly select the first element from all N population members and then select every kth item until we have n items in our sample.

The Advantages of Systematic Sampling

In addition to the ease of drawing samples in a systematic way–for example, by simply measuring distances with a ruler in a file cabinet and sampling every fixed number of inches–the method has some statistical advantages as well. First, when $k = N/n$, the sample estimator of the population mean is unbiased. Second, systematic sampling is usually more precise than simple random sampling because it actually stratifies the population into n strata, each stratum containing k elements. Therefore, systematic sampling is approximately as precise as stratified random sampling with one unit per stratum. The difference between the two methods is that the systematic sample is spread more evenly over the entire population than a stratified sample, because in stratified sampling the samples in the strata are drawn separately. This adds precision in some cases. Systematic sampling is also related to cluster sampling in that it amounts to selecting one cluster out of a population of k clusters.

Estimation of the Population Mean in Systematic Sampling

The systematic sampling estimator of the population mean μ is

$$\bar{X}_{sy} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{16–22}$$

The estimator is, of course, the same as the simple random sampling estimator of the population mean based on a sample of size n. The variance of the estimator in equation 16–22 is difficult to estimate from the results of a single sample. The estimation requires some assumptions about the order of the population. The estimated variances of $\bar{X}_{sy}$ in different situations are given below.

1. When the population values are assumed to be in no particular order with respect to the variable of interest, the estimated variance of the estimator of the mean is the same as in the case of simple random sampling

$$s^2(\bar{X}_{sy}) = \frac{N - n}{Nn} S^2 \tag{16–23}$$

where S^2 is the usual sample variance, and the first term accounts for finite-population correction as well as division by n.

2. When the mean is constant within each stratum of k elements but different from stratum to stratum, the estimated variance of the sample mean is

$$s^2(\overline{X}_{sy}) = \frac{N-n}{Nn} \frac{\sum_{i=1}^{n}(X_i - X_{i+k})^2}{2(n-1)} \qquad (16\text{–}24)$$

3. When the population is assumed to be either increasing or decreasing linearly in the variable of interest, and when the sample size is large, the appropriate estimator of the variance of our estimator of the mean is

$$s^2(\overline{X}_{sy}) = \frac{N-n}{Nn} \frac{\sum_{i=1}^{n}(X_i - 2X_{i+k} + X_{i+2k})^2}{6(n-1)} \qquad (16\text{–}25)$$

for $1 \le i \le n-2$.

There are formulas that apply in more complicated situations as well.

We demonstrate the use of systematic sampling with the following example.

EXAMPLE 16–4

An investor obtains a copy of *The Wall Street Journal* and wants to get a quick estimate of how the New York Stock Exchange has performed since the previous day. The investor knows that there are about 2,100 stocks listed on the NYSE and wants to look at a quick sample of 100 stocks and determine the average price change for the sample. The investor thus decides on an "every 21st" systematic sampling scheme. The investor uses a ruler and finds that this means that a stock should be selected about every 1.5 inches along the listings columns in the *Journal*. The first stock is randomly selected from among the first 21 stocks listed on the NYSE by using a random-number generator in a calculator. The selected stock is the seventh from the top, which happens to be ANR. For the day in question, the price change for ANR is −0.25. The next stock to be included in the sample is the one in position 7 + 21 = 28th from the top. The stock is Aflpb, which on this date had a price change of 0 from the previous day. As mentioned, the selection is not done by counting the stocks, but by the faster method of successively measuring 1.5 inches down the column from each selected stock. The resulting sample of 100 stocks gives a sample mean of $\bar{x}_{sy} = +0.5$ and $S^2 = 0.36$. Give a 95% confidence interval for the average price change of all stocks listed on the NYSE.

Solution

We have absolutely no reason to believe that the order in which the NYSE stocks are listed in *The Wall Street Journal* (i.e., alphabetically) has any relationship to the stocks' price changes. Therefore, the appropriate equation for the estimated variance of $\overline{X}_{sy}$ is equation 16–23. Using this equation, we get

$$s^2(\overline{X}_{sy}) = \frac{N-n}{Nn} S^2 = \frac{2{,}100 - 100}{210{,}000} 0.36 = 0.0034$$

A 95% confidence interval for μ, the average price change on this day for all stocks on the NYSE, is therefore

$$\bar{x}_{sy} \pm 1.96s(\bar{X}_{sy}) = 0.5 \pm (1.96)(\sqrt{0.0034}) = [0.386, 0.614]$$

The investor may be 95% sure that the average stock on the NYSE gained anywhere from \$0.386 to \$0.614.

When sampling for the population proportion, use the same equations as the ones used for simple random sampling if it may be assumed that no inherent order exists in the population. Otherwise use variance estimators given in advanced texts.

The Template

The template for estimating a population mean by systematic sampling is shown in Figure 16–6. The data in the figure correspond to Example 16–4.

FIGURE 16–6 **The Template for Estimating Population Means by Systematic Sampling [Systematic Sampling.xls; Sheet: Sheet 1]**

	A	B	C	D	E	F	G	H	I	J
1	**Systematic Sampling**				Average price change					
2										
3		***N***	***n***	***x*-bar**	***s*²**					
4		2100	100	0.5	0.36					
5										
6				***X*-bar**	0.5					
7				**V(*X*-bar)**	0.003429	Assuming that the population is in no particular order.				
8				**S(*X*-bar)**	0.058554					
9										
10		1–α		**(1-α) CI for X-bar**						
11		95%		**0.5 + or - 0.11476**		or	**0.38524 to 0.61476**			
12										

PROBLEMS

16–10. A tire manufacturer maintains strict quality control of its tire production. This entails frequent sampling of tires from large stocks shipped to retailers. Samples of tires are selected and run continuously until they are worn out, and the average number of miles "driven" in the laboratory is noted. Suppose a warehouse contains 11,000 tires arranged in a certain order. The company wants to select a systematic sample of 50 tires to be tested in the laboratory. Use randomization to determine the first item to be sampled, and give the rule for obtaining the rest of the sample in this case.

16–11. A large discount store gives its sales personnel bonuses based on their average sale amount. Since each salesperson makes hundreds of sales each month, the store management decided to base average sale amount for each salesperson on a random sample of the person's sales. Since records of sales are kept in books, the use of systematic sampling is convenient. Suppose a salesperson has made 855 sales over the month, and management wants to choose a sample of 30 sales for estimation of the average amount of all sales. Suggest a way of doing this so that no problems would result due to the fact that 855 is not an integer multiple of 30. Give

the first element you choose to select, and explain how the rest of the sample is obtained.

16–12. An accountant always audits the second account and every fourth account thereafter when sampling a client's accounts.

a. Does the accountant use systematic sampling? Explain.

b. Explain the problems that may be encountered when this sampling scheme is used.

16–13. Beer sales in a tavern are cyclical over the week, with large volume during weekend nights, lower volume during the beginning of the week, and somewhat higher volume at midweek. Explain the possible problems that could arise, and the conditions under which they might arise, if systematic sampling were used to estimate beer sales volume per night.

16–14. A population is composed of 100 items arranged in some order. Every stratum of 10 items in the order of arrangement tends to be similar in its values. An "every 10th" systematic sample is selected. The first item, randomly chosen, is the 6th item, and its value is 20. The following items in the sample are, of course, the 16th, the 26th, etc. The values of all items in the systematic sample are as follows: 20, 25, 27, 34, 28, 22, 28, 21, 37, 31. Give a 90% confidence interval for the population mean.

16–15. Explain the relationship between the method employed in problem 16–14 and the method of stratified random sampling. Explain the differences between the two methods.

16–6 Nonresponse

Nonresponse to sample surveys is one of the most serious problems that occur in practical applications of sampling methodology. The problem is one of loss of information. For example, suppose that a survey questionnaire dealing with some issue is mailed to a randomly chosen sample of 500 people and that only 300 people respond to the survey. The question is: What can you say about the 200 people who did not respond? This is a very important question, with no immediate answer, precisely because the people did not respond; we know nothing about them. Suppose that the questionnaire asks for a yes or no answer to a particular public issue over which people have differing views, and we want to estimate the proportion of people who would respond yes. People may have such strong views about the issue that those who would respond no may refuse to respond altogether. In this case, the 200 nonrespondents to our survey will contain a higher proportion of "no" answers than the 300 responses we have. But, again, we would not know about this. The result will be a bias. How can we compensate for such a possible bias?

We may want to consider the population as made up of two *strata:* the respondents' stratum and the nonrespondents' stratum. In the original survey, we managed to sample only the respondents' stratum, and this caused the bias. What we need to do is to obtain a random sample from the nonrespondents' stratum. This is easier said than done. Still, there are ways we can at least reduce the bias and get some idea about the proportion of "yes" answers in the nonresponse stratum. This entails *callbacks:* returning to the nonrespondents and asking them again. In some mail questionnaires, it is common to send several requests for response, and these reduce the uncertainty. There may, however, be hard-core refusers who just do not want to answer the questionnaire. These people are likely to have distinct views about the issue in question, and if you leave them out, your conclusions will reflect a significant bias. In such a situation, gathering a small random sample of the hard-core refusers and offering them some monetary reward for their answers may be useful. In cases where people may find the question embarrassing or may worry about revealing their

personal views, a random-response mechanism may be used; here the respondent randomly answers one of two questions, one is the sensitive question, and the other is an innocuous question of no relevance. The interviewer does not know which question any particular respondent answered but does know the probability of answering the sensitive question. This still allows for computation of the aggregated response to the sensitive question while protecting any given respondent's privacy.

16–7 Summary and Review of Terms

In this chapter, we considered some advanced sampling methods that allow for better precision than simple random sampling, or for lowered costs and easier survey implementation. We concentrated on **stratified random sampling,** the most important and useful of the advanced methods and one that offers statistical advantages of improved precision. We then discussed **cluster sampling** and **systematic sampling,** two methods that are used primarily for their ease of implementation and reduced sampling costs. We mentioned a few other advanced methods, which are described in books devoted to sampling methodology. We discussed the problem of **nonresponse.**

ADDITIONAL PROBLEMS

16–16. Bloomingdale's main store in New York has the following departments on its mezzanine level: Stendahl, Ralph Lauren, Beauty Spot, and Lauder Prescriptives. The mezzanine level is managed separately from the other levels, and during the store's postholiday sale, the level manager wanted to estimate the average sales amount per customer throughout the sale. The following table gives the relative weights of the different departments (known from previous operation of the store), as well as the sample means and variances of the different strata for a total sample of 1,000 customers, proportionally allocated to the four strata. Give a 95% confidence interval for the average sale (in dollars) per customer for the entire level over the period of the postholiday sale.

Stratum	Weight	Sample Mean	Sample Variance
Stendahl	0.25	65.00	123.00
Ralph Lauren	0.35	87.00	211.80
Beauty Spot	0.15	52.00	88.85
Lauder Prescriptives	0.25	38.50	100.40

Note: We assume that shoppers visit the mezzanine level to purchase from only one of its departments. Since the brands and designers are competitors, and since shoppers are known to have a strong brand loyalty in this market, the assumption seems reasonable.

16–17. A state department of transportation is interested in sampling commuters to determine certain of their characteristics. The department arranges for its field workers to board buses at random as well as stop private vehicles at intersections and ask the commuters to fill out a short questionnaire. Is this method cluster sampling? Explain.

16–18. Use systematic sampling to estimate the average performance of all stocks in one of the listed stock exchanges on a given day. Compare your results with those reported in the media for the day in question.

16–19. An economist wants to estimate average annual profits for all businesses in a given community and proposes to draw a systematic sample of all businesses listed

in the local Yellow Pages. Comment on the proposed methodology. What potential problem do you foresee?

16–20. In an "every 23rd" systematic sampling scheme, the first item was randomly chosen to be the 17th element. Give the numbers of 6 sample items out of a population of 120.

16–21. A quality control sampling scheme was carried out by Sony for estimating the percentage of defective radios in a large shipment of 1,000 containers with 100 radios in each container. Twelve containers were chosen at random, and every radio in them was checked. The numbers of defective radios in each of the containers are 8, 10, 4, 3, 11, 6, 9, 10, 2, 7, 6, 12. Give a 95% confidence interval for the proportion of defective radios in the entire shipment.

16–22. Suppose that the radios in the 1,000 containers of problem 16–21 were produced in five different factories, each factory known to have different internal production controls. Each container is marked with a number denoting the factory where the radios were made. Suggest an appropriate sampling method in this case, and discuss its advantages.

16–23. The makers of Taster's Choice instant coffee want to estimate the proportion of underfilled jars of a given size. The jars are in 14 warehouses around the country, and each warehouse contains crates of cases of jars of coffee. Suggest a sampling method, and discuss it.

16–24. Cadbury, Inc., is interested in estimating people's responses to a new chocolate. The company believes that people in different age groups differ in their preferences for chocolate. The company believes that in the region of interest, 25% of the population are children, 55% are young adults, and 20% are older people. A proportional allocation of a total random sample of size 1,000 is undertaken, and people's responses on a scale of 0 to 100 are solicited. The results are as follows. For the children, $\bar{x} = 90$ and $s = 5$; for the young adults, $\bar{x} = 82$ and $s = 11$; and for the older people, $\bar{x} = 88$ and $s = 6$. Give a 95% confidence interval for the population average rating for the new chocolate.

16–25. For problem 16–24, suppose that it costs twice as much money to sample a child as the younger and older adults, where costs are the same per sampled person. Use the information in problem 16–24 (the weights and standard deviations) to determine an optimal allocation of the total sample.

16–26. Refer to the situation in problem 16–24. Suppose that the following relative age frequencies in the population are known:

Age Group	Frequency
Under 10	0.10
10 to 15	0.10
16 to 18	0.05
19 to 22	0.05
23 to 25	0.15
26 to 30	0.15
31 to 35	0.10
36 to 40	0.10
41 to 45	0.05
46 to 50	0.05
51 to 55	0.05
56 and over	0.05

Define strata to be used in the survey.

16–27. Name two sampling methods that are useful when there is information about a variable related to the variable of interest.

16–28. Suppose that a study was undertaken using a simple random sample from a particular population. When the results of the study became available, they revealed that the population could be viewed as consisting of several strata. What can be done now?

16–29. For problem 16–28, suppose that the population is viewed as comprising 18 strata. Is using this number of strata advantageous? Are there any alternative solutions?

16–30. Discuss and compare the three sampling methods: cluster sampling, stratified sampling, and systematic sampling.

16–31. The following table reports return on capital for insurance companies. Consider the data a population of U.S. insurance companies, and select a random sample of firms to estimate mean return on capital. Do the sampling two ways: first, take a systematic sample considering the entire list as a uniform, ordered population; and second, use stratified random sampling, the strata being the types of insurance company. Compare your results.

Company Diversified	Return on Capital Latest 12 Mos. %	Company Life & Health	Return on Capital Latest 12 Mos. %	Company Property & Casualty	Return on Capital Latest 12 Mos. %
Marsh & McLennan Cos	25.4	Conseco	13.7	20th Century Industries	25.1
Loews	13.8	First Capital Holding	10.7	Geico	20.4
American Intl Group	14.6	Torchmark	18.4	Argonaut Group	17.1
General Rental	17.4	Capital Holding	9.9	Hartford Steam Boiler	19.1
Safeco	11.6	American Family	8.5	Progressive	10.1
Leucadia National	27.0	Kentucky Central Life	6.3	WR Berkley	11.5
CNA Financial	8.6	Provident Life & Acc	13.1	Mercury General	28.3
Aon	12.8	NWNL	8.4	Selective Insurance	13.6
Kemper	1.4	UNUM	13.2	Hanover Insurance	6.3
Cincinnati Financial	10.7	Liberty Corp	10.1	St Paul Cos	16.9
Reliance Group	23.1	Jefferson-Pilot	9.5	Chubb	14.3
Alexander & Alexander	9.9	USLife	6.7	Ohio Casualty	9.3
Zenith National Ins	9.8	American Natl Ins	5.2	First American Finl	4.8
Old Republic Intl	13.4	Monarch Capital	0	Berkshire Hathaway	7.3
Transamerica	7.5	Washington National	1.2	ITT	10.2
Uslico	7.7	Broad	8.0	USF&G	6.6
Aetna Life & Cas	8.0	First Executive	0	Xerox	5.3
American General	8.2	ICH	0	Orion Capital	10.8
Lincoln National	9.2			Fremont General	12.6
Sears, Roebuck	7.2			Foremost Corp of Amer	0
Independent Insurance	7.8			Continental Corp	6.6
Cigna	7.0			Alleghany	9.0
Travelers	0				
American Bankers	8.3				
Unitrin	6.1				

From "Annual Report on American Industry," *Forbes,* January 7, 1991. Reprinted by permission of Forbes Magazine. © 2001 Forbes Inc.

CASE 21 The Boston Redevelopment Authority

The Boston Redevelopment Authority is mandated with the task of improving and developing urban areas in Boston. One of the Authority's main concerns is the development of the community of Roxbury. This community has undergone many changes in recent years, and much interest is given to its future development.

Currently, only 2% of the total land in this community is used in industry, and 9% is used commercially. As part of its efforts to develop the community, the Boston Redevelopment Authority is interested in determining the attitudes of the residents of Roxbury toward the development of more business and industry in their region. The authority therefore plans to sample residents of the community to determine their views and use the sample or samples to infer about the views of all residents of the community. Roxbury is divided into 11 planning subdistricts. The population density is believed to be uniform across all 11 subdistricts, and the population of each subdistrict is approximately proportional to the subdistrict's size. There is no known list of all the people in Roxbury. A map of the community is shown in Exhibit 1. Advise the Boston Redevelopment Authority on designing the survey.

EXHIBIT 1 **Roxbury Planning Subdistrict**