• *Readable* • *Modern* • *Practical* • *Complete*

Sharon L. Lohr's *Sampling: Design and Analysis* brings the field of survey sampling into the real world, with true-life examples drawn from a variety of disciplines and surveys. With an emphasis on evaluating real surveys and solving real problems, this book is equal parts instructional tool and reference guide—as useful for the reader new to the subject as it is for the survey practitioner. A disk—complete with data sets and the SURVEY computer program—provides additional support for the numerous practical exercises, graphs, projects, and other useful features of this thoroughly modern book.

"This is the sampling text that I have been waiting for; it is the book I wish I could have written. There has been a real need for a text that covers theoretical material at this level and also includes real applications and modern topics in sampling, such as the model-based approach . . . Very readable . . . A very nice balance between traditional, classical survey sampling topics and theory, and more modern topics and concerns."

Elizabeth A. Stasny
The Ohio State University, Columbus

THIS BOOK IS PRINTED ON ACID-FREE RECYCLED PAPER

www.duxbury.com

JOIN US ON THE INTERNET
WWW: http://www.thomson.com
E-MAIL: findit@kiosk.thomson.com

VISIT DUXBURY ON THE INTERNET
For information on our products and services, to download demos, and to search our catalogs, point your web browser to
http://www.duxbury.com
or e-mail us: **duxbury_info@wadsworth.com**

ISBN 0-534-35361-4
90000
9 780534 353612

A service of ITP®

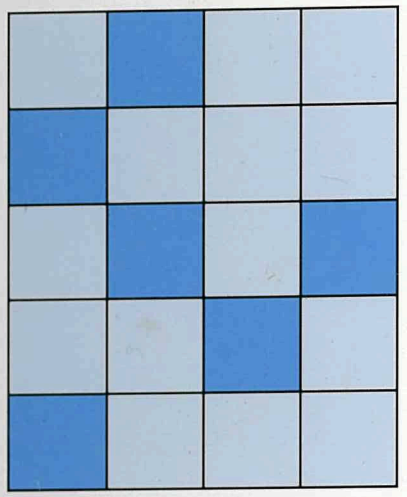
DUXBURY PRESS

ITP

Lohr

Sampling: Design and Analysis

Duxbury



Sampling: Design and Analysis

Sharon L. Lohr

DUXBURY PRESS

UNIVERSITY BOOKSHOP
CANTERBURY

4

Stratified Sampling

One of the things she [Mama] taught me should be obvious to everyone, but I still find a lot of cooks who haven't figured it out yet. Put the food on first that takes the longest to cook.

—Pearl Bailey, *Pearl's Kitchen*

4.1 What Is Stratified Sampling?

Often, we have supplementary information that can help us design our sample. For example, we would know before undertaking an income survey that men generally earn more than women, that New York City residents pay more for housing than residents of Des Moines, or that rural residents shop for groceries less frequently than urban residents.

If the variable we are interested in takes on different mean values in different subpopulations, we may be able to obtain more precise estimates of population quantities by taking a **stratified random sample.** The word *stratify* comes from Latin words meaning "to make layers"; we divide the population into H subpopulations, called **strata.** The strata do not overlap, and they constitute the whole population so that each sampling unit belongs to exactly one stratum. We draw an independent probability sample from each stratum, then pool the information to obtain overall population estimates.

We use stratified sampling for one or more of the following reasons:

1 We want to be protected from the possibility of obtaining a really bad sample. When taking a simple random sample (SRS) of size 100 from a population of 1000 male and 1000 female students, obtaining a sample with no or very few males is theoretically possible, although such a sample is not likely to occur. Most people would not consider such a sample to be representative of the population and would worry that men and women might respond differently on the item of interest. In a stratified sample, one could take an SRS of 50 males and an independent SRS of 50 females, guaranteeing that the proportion of males in the sample is the same as that in the population. With this design, a sample with no or few males cannot be selected.

2 We may want data of known precision for subgroups. These subgroups should be the strata, which then coincide with the domains of study. McIlwee and Robinson (1992) sampled graduates from electrical and mechanical engineering programs at public universities in southern California. They were interested in comparing the educational and workforce experiences of male and female graduates, so they stratified their sampling frame by gender and took separate random samples of male graduates and female graduates. Because there were many more male than female graduates, they sampled a higher fraction of female graduates than male graduates in order to obtain comparable precisions for the two groups.

3 A stratified sample may be more convenient to administer and may result in a lower cost for the survey. For example, different sampling approaches may be used for different strata. In a survey of businesses, a mail survey might be used for large firms, whereas a personal or telephone interview is used for small firms. In other surveys, different sampling methods may be needed in urban and rural strata.

4 Stratified sampling, if done correctly, will give more precise (having lower variance) estimates for the whole population. Persons of different ages tend to have different blood pressures, so in a blood pressure study it would be helpful to stratify by age groups. If studying the concentration of plants in an area, one would stratify by type of terrain; marshes would have different plants than woodlands. Stratification works for lowering the variance because the variance within each stratum is often lower than the variance in the whole population. Prior knowledge can be used to save money in the sampling procedure.

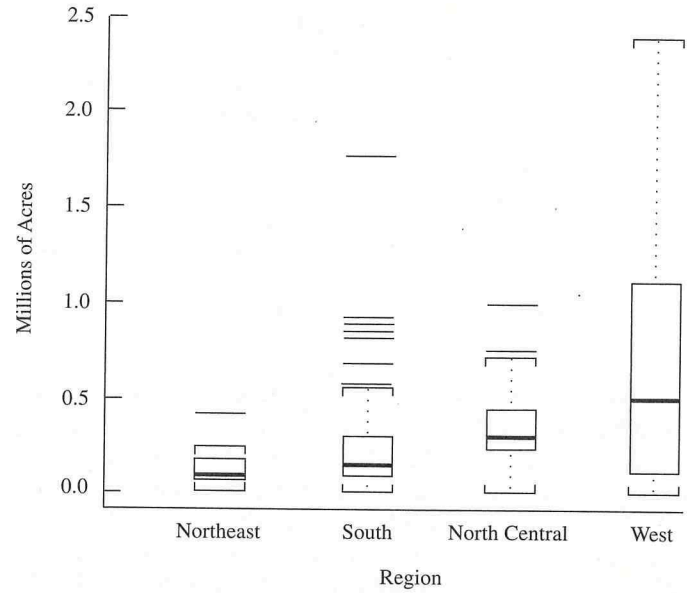
EXAMPLE 4.1 Refer to Example 2.4, in which we took an SRS to estimate the average number of farm acres per county. In Example 2.4, we noted that, even though we scrupulously generated a random sample, some areas were overrepresented and others not represented at all. Taking a stratified sample can provide some balance in the sample on the stratifying variable.

The SRS in Example 2.4 exhibited a wide range of values for y_i, the number of acres devoted to farms in county i in 1992. You might conjecture that part of the large variability arises because counties in the western United States are larger, and thus tend to have larger values of y, than counties in the eastern United States.

For this example, we use the four census regions of the United States—Northeast, North Central, South, and West—as strata. The SRS in Example 2.4 sampled about 10% of the population; to compare the results of the stratified sample with the SRS, we also sample about 10% of the counties in each stratum. (We discuss other stratified sampling designs later in the chapter.)

Stratum	Number of Counties in Stratum	Number of Counties in Sample
Northeast	220	21
North Central	1054	103
South	1382	135
West	422	41
Total	3078	300

FIGURE 4.1
The boxplot of data from Example 4.1. The thick line for each region is the median of the sample data from that region; the other horizontal lines in the boxes are the 25th and 75th percentiles. The Northeast region has a relatively low median and small variance; the West region, however, has a much higher median and variance. The distribution of farm acreage appears to be positively skewed in each of the regions.



We select four separate SRSs, one from each of the four strata. To select the SRS from the Northeast stratum, we number the counties in that stratum from 1 to 220 and select 21 numbers randomly from $\{1, \ldots, 220\}$. We follow a similar procedure for the other three strata, selecting 103 counties at random from the 1054 in the North Central region, 135 counties from the 1382 in the South, and 41 counties from the 422 in the West. The four SRSs are independent: Knowing which counties are in the sample from the Northeast tells us nothing about which counties are in the sample from the South.

The data sampled from all four strata are in data file agstrat.dat. A boxplot, showing the data for each stratum, is in Figure 4.1. Summary statistics for each stratum are given below:

Region	Sample Size	Average	Variance
Northeast	21	97,629.8	7,647,472,708
North Central	103	300,504.2	29,618,183,543
South	135	211,315.0	53,587,487,856
West	41	662,295.5	396,185,950,266

Since we took an SRS in each stratum, we can use Equations (2.12) and (2.14) to estimate the population quantities for each stratum. We use

$$(220)(97{,}629.81) = 21{,}478{,}558.2$$

to estimate the total number of acres devoted to farms in the Northeast, with estimated variance

$$(220)^2\left(1-\frac{21}{220}\right)\frac{7{,}647{,}472{,}708}{21} = 1.594316 \times 10^{13}.$$

The following table gives estimates of the total number of farm acres and estimated variance of the total for each of the four strata:

Stratum	Estimated Total of Farm Acres	Estimated Variance of Total
Northeast	21,478,558	1.59432×10^{13}
North Central	316,731,379	2.88232×10^{14}
South	292,037,391	6.84076×10^{14}
West	279,488,706	1.55365×10^{15}
Total	909,736,034	2.5419×10^{15}

We can estimate the total number of acres devoted to farming in the United States by adding the totals for each stratum; as sampling was done independently in each stratum, the variance of the U.S. total is the sum of the variances of the population stratum totals. Thus, we estimate the total number of acres devoted to farming as 909,736,034, with standard error $\sqrt{2.5419 \times 10^{15}} = 50{,}417{,}248$. We would estimate the average number of acres devoted to farming per county as $909{,}736{,}034/3078 = 295{,}560.7649$, with standard error $50{,}417{,}248/3078 = 16{,}379.87$.

For comparison, the estimate of the total in Example 2.4, using an SRS of size 300, was 916,927,110, with standard error 58,169,381. For this example, stratified sampling ensures that each region of the United States is represented in the sample and produces an estimate with a slightly smaller standard error than an SRS with the same number of observations. The sample variance in Example 2.4 was $s^2 = 1.1872 \times 10^{11}$. Only the West had sample variance larger than s^2; the sample variance in the Northeast was only 7.647×10^9.

Observations within many strata tend to be more homogeneous than observations in the population as a whole, and the reduction in variance in the individual strata often leads to a reduced variance for the population estimate. In this example, the relative gain from stratification can be estimated by the ratio

$$\frac{\text{estimated variance from stratification, with } n = 300}{\text{estimated variance from SRS, with } n = 300} = \frac{2.5419 \times 10^{15}}{3.3837 \times 10^{15}} = 0.75.$$

If these figures were the population variances, we would expect that we would need only $(300)(0.75) = 225$ observations with a stratified sample to obtain the same precision as from an SRS of 300 observations.

Of course, no law says that you must sample the same fraction of observations in every stratum. In this example, there is far more variability from county to county in the western region; if acres devoted to farming were the primary variable of interest, you would reduce the variance of the estimated total even further by taking a higher

sampling fraction in the western region than in the other regions. You will explore an alternative sampling design in Exercise 12. ■

4.2 Theory of Stratified Sampling

We divide the population of N sampling units into H "layers," or strata, with N_h sampling units in the hth stratum. For stratified sampling to work, we must know the values of $N_1, N_2, \ldots, N_H$ and must have

$$N_1 + N_2 + \cdots + N_H = N,$$

where N is the total number of units in the entire population.

In **stratified random sampling,** the simplest form of stratified sampling, we independently take an SRS from each stratum so that n_h observations are randomly selected from the population units in stratum h. Define $\mathcal{S}_h$ to be the set of n_h units in the SRS for stratum h.

Notation for Stratification The population quantities are:

$$y_{hj} = \text{value of } j\text{th unit in stratum } h$$

$$t_h = \sum_{j=1}^{N_h} y_{hj} = \text{population total in stratum } h$$

$$t = \sum_{h=1}^{H} t_h = \text{population total}$$

$$\bar{y}_{hU} = \frac{\sum_{j=1}^{N_h} y_{hj}}{N_h} = \text{population mean in stratum } h$$

$$\bar{y}_U = \frac{t}{N} = \frac{\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj}}{N} = \text{overall population mean}$$

$$S_h^2 = \sum_{j=1}^{N_h} \frac{(y_{hj} - \bar{y}_{hU})^2}{N_h - 1} = \text{population variance in stratum } h$$

Corresponding quantities for the sample, using SRS estimates within each stratum, are:

$$\bar{y}_h = \frac{\sum_{j \in \mathcal{S}_h} y_{hj}}{n_h}$$

$$\hat{t}_h = \frac{N_h}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj} = N_h \bar{y}_h$$

$$s_h^2 = \sum_{j \in \mathcal{S}_h} \frac{(y_{hj} - \bar{y}_h)^2}{n_h - 1}$$

Suppose we only sampled the hth stratum. In effect, we have a population of N_h units and take an SRS of n_h units. Then we would estimate $\bar{y}_{hU}$ by $\bar{y}_h$, and t_h by $\hat{t}_h = N_h\bar{y}_h$. The population total is $t = \sum_{h=1}^{H} t_h$, so we estimate t by

$$\hat{t}_{\text{str}} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h\bar{y}_h. \tag{4.1}$$

To estimate $\bar{y}_U$, then, use

$$\bar{y}_{\text{str}} = \frac{\hat{t}_{\text{str}}}{N} = \sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_h. \tag{4.2}$$

This is a weighted average of the sample stratum averages; the weights are the relative sizes of the strata. To use stratified sampling, the sizes or relative sizes of the strata must be known.

The properties of these estimators follow directly from the properties of SRS estimators:

- **Unbiasedness.** $\bar{y}_{\text{str}}$ and $\hat{t}_{\text{str}}$ are unbiased estimators of $\bar{y}_U$ and t. This is true because

$$E\left[\sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_h\right] = \sum_{h=1}^{H} \frac{N_h}{N} E[\bar{y}_h] = \sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_{hU} = \bar{y}_U.$$

- **Variance of the estimators.** Since we are sampling independently from the strata and we know $V(\hat{t}_h)$ from SRS theory, the properties of expected value (p. 427) and Equation (2.13) imply that

$$V(\hat{t}_{\text{str}}) = \sum_{h=1}^{H} V(\hat{t}_h) = \sum_{h=1}^{H}\left(1 - \frac{n_h}{N_h}\right)N_h^2\frac{S_h^2}{n_h}. \tag{4.3}$$

- **Variance estimates for stratified samples.** We can obtain an unbiased estimator of $V(\hat{t}_{\text{str}})$ by substituting the sample estimates s_h^2 for the population quantities S_h^2. Note that, to estimate the variances, we need to sample at least two units from each stratum:

$$\hat{V}(\hat{t}_{\text{str}}) = \sum_{h=1}^{H}\left(1 - \frac{n_h}{N_h}\right)N_h^2\frac{s_h^2}{n_h}, \tag{4.4}$$

$$\hat{V}(\bar{y}_{\text{str}}) = \frac{1}{N^2}\hat{V}(\hat{t}_{\text{str}}) = \sum_{h=1}^{H}\left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2\frac{s_h^2}{n_h}. \tag{4.5}$$

As always, the standard error of an estimator is the square root of the estimated variance: $\text{SE}(\bar{y}_{\text{str}}) = \sqrt{\hat{V}(\bar{y}_{\text{str}})}$.

- **Confidence intervals for stratified samples.** If either (1) the sample sizes within each stratum are large or (2) the sampling design has a large number of strata, an approximate $100(1-\alpha)\%$ confidence interval (CI) for the mean is

$$\bar{y}_{\text{str}} \pm z_{\alpha/2}\,\text{SE}(\bar{y}_{\text{str}}).$$

The central limit theorem used for constructing this confidence interval is stated in Krewski and Rao (1981). Some survey researchers use the percentile of a

t distribution with $n - H$ degrees of freedom (df) rather than the percentile of the normal distribution.

EXAMPLE 4.2 Siniff and Skoog (1964) used stratified random sampling to estimate the size of the Nelchina herd of Alaskan caribou in February 1962. In January and early February, several sampling techniques were field-tested. The field tests told the investigators that several of the proposed sampling units, such as equal-flying-time sampling units, were difficult to implement in practice and that an equal-area sampling unit of 4 square miles (mi^2) would work well for the survey. The biologists used preliminary estimates of caribou densities to divide the area of interest into six strata; each stratum was then divided into a grid of 4-mi^2 sampling units. Stratum A, for example, contained $N_1 = 400$ sampling units; $n_1 = 98$ of these were randomly selected to be in the survey. The following data were reported:

Stratum	N_h	n_h	$\bar{y}_h$	s_h^2
A	400	98	24.1	5,575
B	30	10	25.6	4,064
C	61	37	267.6	347,556
D	18	6	179.0	22,798
E	70	39	293.7	123,578
F	120	21	33.2	9,795

With the data in this form, using a spreadsheet to do the calculations necessary for stratified sampling is easy. The spreadsheet shown in Table 4.1 simplifies the calculations that the estimated total number of caribou is 54,497 with standard error 5840. An approximate 95% CI for the total number of caribou is

$$54{,}497 \pm 1.96(5840) = [43{,}051,\ 65{,}943].$$

TABLE 4.1
Spreadsheet for Calculations in Example 4.2

	A	B	C	D	E	F	G
1	Stratum	N_h	n_h	$\bar{y}_h$	s_h^2	$\hat{t}_h = N_h\bar{y}_h$	$\left(1 - \frac{n_h}{N_h}\right)N_h^2\frac{s_h^2}{n_h}$
2	A	400	98	24.1	5,575	9,640	6,872,040.82
3	B	30	10	25.6	4,064	768	243,840.00
4	C	61	37	267.6	347,556	16,324	13,751,945.51
5	D	18	6	179.0	22,798	3,222	820,728.00
6	E	70	39	293.7	123,578	20,559	6,876,006.67
7	F	120	21	33.2	9,795	3,984	5,541,171.43
8	total		211			54,497	34,105,732.43
9	sqrt(total)						5,840.01

Of course, this confidence interval only reflects the uncertainty due to sampling error; if the field procedure for counting caribou tends to miss animals, then the entire confidence interval will be too low. ■

Stratified Sampling for Proportions As we observed in Section 2.3, a proportion is a mean of a variable that takes on values 0 and 1. To make inferences about proportions, we just use Equations (4.1)–(4.5), with $\bar{y}_h = \hat{p}_h$ and $s_h^2 = [n_h/(n_h - 1)]\hat{p}_h(1 - \hat{p}_h)$. Then,

$$\hat{p}_{\text{str}} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h \tag{4.6}$$

and

$$\hat{V}(\hat{p}_{\text{str}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}. \tag{4.7}$$

Estimating the total number of population units having a specified characteristic is similar:

$$\hat{t}_{\text{str}} = \sum_{h=1}^{H} N_h \hat{p}_h.$$

Thus, the estimated total number of population units with the characteristic is the sum of the estimated totals in each stratum. Similarly, $\hat{V}(\hat{t}_{\text{str}}) = N^2 \hat{V}(\hat{p}_{\text{str}})$.

EXAMPLE 4.3 The American Council of Learned Societies (ACLS) used a stratified random sample of selected ACLS societies in seven disciplines to study publication patterns and computer and library use among scholars who belong to one of the member organizations of the ACLS (Morton and Price 1989). The data are shown in Table 4.2.

Ignoring the nonresponse for now (we'll return to the nonresponse in Exercise 9 in Chapter 8) and supposing there are no duplicate memberships, let's use the stratified sample to estimate the percentage and number of respondents of the major societies in those seven disciplines who are women. Here, let N_h be the membership figures

TABLE 4.2
Data from ACLS Survey

Discipline	Membership	Number Mailed	Valid Returns	Female Members (%)
Literature	9,100	915	636	38
Classics	1,950	633	451	27
Philosophy	5,500	658	481	18
History	10,850	855	611	19
Linguistics	2,100	667	493	36
Political science	5,500	833	575	13
Sociology	9,000	824	588	26
Totals	44,000	5,385	3,835	

and let n_h be the number of valid surveys. Thus,

$$\hat{p}_{\text{str}} = \sum_{h=1}^{7} \frac{N_h}{N} \hat{p}_h = \frac{9100}{44{,}000}0.38 + \cdots + \frac{9000}{44{,}000}0.26 = 0.2465$$

and

$$\text{SE}(\hat{p}_{\text{str}}) = \sqrt{\sum_{h=1}^{7} \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}} = 0.0071.$$

The estimated total number of female members in the societies is $\hat{t}_{\text{str}} = 44{,}000 \times (0.2465) = 10{,}847$, with $\text{SE}(\hat{t}_{\text{str}}) = 44{,}000(.0071) = 312$. ■

4.3 Sampling Weights

The stratified sampling estimator $\hat{t}_{\text{str}}$ can be expressed as a weighted sum of the individual sampling units. Using (4.1),

$$\hat{t}_{\text{str}} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} \frac{N_h}{n_h} y_{hj}.$$

The **sampling weight** $w_{hj} = (N_h/n_h)$ can be thought of as the number of units in the population represented by the sample member (h, j). If the population has 1600 men and 400 women and the stratified sample design specifies sampling 200 men and 200 women, then each man in the sample has weight 8 and each woman has weight 2. Each woman in the sample represents herself and 1 other woman not selected to be in the sample, and each man represents himself and 7 other men not in the sample. Note that the probability of selecting the jth unit in the hth stratum to be in the sample is $\pi_{hj} = n_h/N_h$, the sampling fraction in the hth stratum. Thus, the sampling weight is simply the reciprocal of the probability of selection:

$$w_{hj} = \frac{1}{\pi_{hj}}. \tag{4.8}$$

The sum of the sampling weights equals the population size N; each sampled unit "represents" a certain number of units in the population, so the whole sample "represents" the whole population. This identity provides a check on whether you have constructed your weight variable correctly: If the sum of the weights for your sample is something other than N, then you have made a mistake somewhere.

The stratified estimate of the population total may thus be written as

$$\hat{t}_{\text{str}} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}, \tag{4.9}$$

and the estimate of the population mean as

$$\bar{y}_{\text{str}} = \frac{\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj}}. \tag{4.10}$$

EXAMPLE 4.4 For the caribou survey in Example 4.2, the weights are

Stratum	N_h	n_h	w_{hj}
A	400	98	4.08
B	30	10	3.00
C	61	37	1.65
D	18	6	3.00
E	70	39	1.79
F	120	21	5.71

In stratum A, each sampling unit of 4 mi^2 represents 4.08 sampling units in the stratum (including itself); in stratum B, a sampling unit in the sample represents itself and 2 other sampling units that are not in the sample. To estimate the population total, then, a new variable of weights could be constructed. This variable would contain the value 4.08 for every observation in stratum A, 3.00 for every observation in stratum B, and so on. ■

EXAMPLE 4.5 The sample in Example 4.1 was designed so that each county in the United States would have approximately the same probability of appearing in the sample. To estimate the total number of acres devoted to agriculture in the United States, we can create a column in the data set (column 17 in the file agstrat.dat) consisting of the sampling weights. The weight column contains the value 220/21 for counties in the Northeast stratum, 1054/103 for the North Central counties, 1382/135 for the South counties, and 422/41 for the West counties. We can use (4.9) to estimate the population total by forming a new column containing the product of the variables *weight* and *acres92*, then calculating the sum of the new column. In doing so, we calculate $\hat{t}_{\text{str}} = 909{,}736{,}035$, the same estimate (except for roundoff error) as obtained in Example 4.1.

The variable *weight* in column 17 can be used to estimate the population total for every variable measured in the sample. Note, however, that you cannot calculate the standard error of $\hat{t}_{\text{str}}$ unless you know the stratification. Equation (4.4) requires that you calculate the variance separately within each stratum; the weights do not tell you the stratum membership of the observations. ■

4.4 Allocating Observations to Strata

So far we have simply analyzed data from a survey that someone else has designed. Designing the survey is the most important part of using a survey in research: If the survey is badly designed, then no amount of analysis will yield the needed information. In this section, different methods of allocating observations to strata are discussed.

4.4.1 Proportional Allocation

If you are taking a stratified sample to ensure that the sample reflects the population with respect to the stratification variable and you would like your sample to be a miniature version of the population, you should use proportional allocation when designing the sample.

In **proportional allocation,** so called because the number of sampled units in each stratum is proportional to the size of the stratum, the probability of selection $\pi_{hj} = n_h/N_h$ is the same ($= n/N$) for all strata; in a population of 2400 men and 1600 women, proportional allocation with a 10% sample would mean sampling 240 men and 160 women. Thus, the probability that an individual will be selected to be in the sample, n/N, is the same as in an SRS, but many of the "bad" samples that could occur in an SRS (for example, a sample in which all 400 persons are men) cannot be selected in a stratified sample with proportional allocation.

If proportional allocation is used, each unit in the sample represents the same number of units in the population: In our example, each man in the sample represents 10 men in the population, and each woman represents 10 women in the population. The sampling weight for every unit in the sample thus equals 10, and the stratified sampling estimate of the population mean is simply the average of all observations. When every unit in the sample has the same weight and represents the same number of units in the population, the sample is called **self-weighting.** The sample in Example 4.1 was designed to be self-weighting. In a self-weighting sample, $\bar{y}_{\text{str}}$ is the average of all observations in the sample.

When the strata are large enough, the population variance of $\bar{y}_{\text{str}}$ under proportional allocation is usually at most as large as the population variance of $\bar{y}$, using the same number of observations but collected in a random sample. This is true no matter how silly the stratification scheme may be. To see why this might be so, let's display the between-strata and within-strata variances, for proportional allocation, in an ANOVA table for the population (Table 4.3).

In a stratified sample of size n with proportional allocation, since $n_h/N_h = n/N$, Equation (4.3) implies that

$$\begin{aligned} V_{\text{prop}}(\hat{t}_{\text{str}}) &= \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)N_h^2\frac{S_h^2}{n_h} \\ &= \left(1-\frac{n}{N}\right)\frac{N}{n}\sum_{h=1}^{H}N_hS_h^2 \\ &= \left(1-\frac{n}{N}\right)\frac{N}{n}\left(\text{SSW}+\sum_{h=1}^{H}S_h^2\right). \end{aligned}$$

TABLE 4.3
Population ANOVA Table

Source	df	Sum of Squares
Between strata	$H-1$	$\text{SSB} = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(\bar{y}_{hU}-\bar{y}_U)^2 = \sum_{h=1}^{H}N_h(\bar{y}_{hU}-\bar{y}_U)^2$
Within strata	$N-H$	$\text{SSW} = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\bar{y}_{hU})^2 = \sum_{h=1}^{H}(N_h-1)S_h^2$
Total, about $\bar{y}_U$	$N-1$	$\text{SSTO} = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\bar{y}_U)^2 = (N-1)S^2$

The sums of squares add up, with SSTO = SSB + SSW, so

$$\begin{aligned} V_{\text{SRS}}(\hat{t}) &= \left(1 - \frac{n}{N}\right) N^2 \frac{S^2}{n} \\ &= \left(1 - \frac{n}{N}\right) \frac{N^2}{n} \frac{\text{SSTO}}{N-1} \\ &= \left(1 - \frac{n}{N}\right) \frac{N^2}{n(N-1)} (\text{SSW} + \text{SSB}) \\ &= V_{\text{prop}}(\hat{t}_{\text{str}}) + \left(1 - \frac{n}{N}\right) \frac{N}{n(N-1)} \left[N(\text{SSB}) - \sum_{h=1}^{H} (N - N_h) S_h^2 \right]. \end{aligned}$$

This result shows us that proportional allocation with stratification always gives an equal or smaller variance than SRS *unless*

$$\text{SSB} < \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) S_h^2. \tag{4.11}$$

This rarely happens when the N_h's are large; generally, the large population sizes of the strata will force $N_h(\bar{y}_{hU} - \bar{y}_U)^2 > S_h^2$. In general, the variance of the estimator of t from proportional allocation will be smaller than the variance of the estimator of t from simple random sampling. The more unequal the stratum means $\bar{y}_{hU}$, the more precision you will gain by using proportional allocation. Of course, this result only holds for population variances; it is possible for a variance estimate from proportional allocation to be larger than that from an SRS merely because the sample selected resulted in a large sample variance.

4.4.2 Optimal Allocation

If the variances S_h^2 are more or less equal across all the strata, proportional allocation is probably the best allocation for increasing precision. In cases where the S_h^2's vary greatly, **optimal allocation** can result in smaller costs. In practice, when we are sampling units of different sizes, the larger units are likely to be more variable than the smaller units, and we would sample them at a higher rate. For example, if we were to take a sample of American corporations and our goal was to estimate the amount of trade with Europe, the variation among the larger corporations would be greater than the variation among smaller ones. As a result, we would sample a higher percentage of the larger corporations. Optimal allocation works well for sampling units such as corporations, cities, and hospitals, which vary greatly in size. It is also effective when some strata are much more expensive to sample than others.

Neter (1978) tells of a study done by the Chesapeake and Ohio (C&O) Railroad Company to determine how much revenue they should get from interline freight shipments, since the total freight from a shipment that traveled along several railroads was divided among the different railroads. The C&O took a stratified sample of waybills, the documents that detailed the goods, route, and charges for the shipments. The waybills were stratified by the total freight charges, and all waybills with charges of over $40 were sampled, whereas only 1% of the waybills with charges less than $5 were sampled. The justification was that there was little variability among the amounts due the C&O in the stratum of the smallest total freight charges, whereas the variability in the stratum with charges of over $40 was much higher.

EXAMPLE 4.6 How are musicians paid when their compositions are performed? In the United States, many composers are affiliated with the American Society of Composers, Authors, and Publishers (ASCAP). Television networks, local television and radio stations, services such as Muzak, symphony orchestras, restaurants, nightclubs, and other operations pay ASCAP an annual license fee, based largely on the size of the audience, that allows them to play compositions in the ASCAP catalog. ASCAP then distributes royalties to composers whose works are played.

Theoretically, an ASCAP member should get royalties every time one of his or her compositions is played. Taking a census of every piece of music played in the United States, however, would be impractical; to estimate the amount of royalties due to members, ASCAP uses sampling. According to Dobishinski (1991) and "The ASCAP Advantage" (1992), ASCAP relies on television producers' cue sheets, which provide details on the music used in a program, to identify and tabulate musical pieces played on network television and major cable channels. About 60,000 hours of tapes are made from radio broadcasts each year, and experts identify the musical compositions aired in these broadcasts.

Stratified sampling is used to sample radio stations for the survey. Radio stations are grouped into strata based on the license fee paid to ASCAP, the type of community the station is in, and the geographic region. As stations paying higher license fees contribute more money for royalties, they are more likely to be sampled; once in the sample, high-fee stations are taped more often than low-fee stations. ASCAP thus uses a form of optimal allocation in taping: Strata with the highest radio fees, and thus with the highest variability in royalty amounts, have larger sampling fractions than strata containing radio stations that pay small fees. ■

The objective in sampling is to gain the most information for the least cost. A simple cost function is given below: Let C represent total cost, c_0 represent overhead costs such as maintaining an office, and c_h represent the cost of taking an observation in stratum h so that

$$C = c_0 + \sum_{h=1}^{H} c_h n_h. \tag{4.12}$$

We want to allocate observations to strata in order to minimize $V(\bar{y}_{\text{str}})$ for a given total cost C or, equivalently, to minimize C for a fixed $V(\bar{y}_{\text{str}})$. Suppose the costs $c_1, c_2, \ldots, c_H$ are fixed. To minimize the variance for a fixed cost, we can prove, using calculus, that the optimal allocation has n_h proportional to

$$\frac{N_h S_h}{\sqrt{c_h}} \tag{4.13}$$

for each h (see Exercise 22). Thus, the optimal sample size in stratum h is

$$n_h = \left(\frac{\dfrac{N_h S_h}{\sqrt{c_h}}}{\displaystyle\sum_{l=1}^{H} \frac{N_l S_l}{\sqrt{c_l}}} \right) n.$$

We then sample heavily within a stratum if

- The stratum accounts for a large part of the population.
- The variance within the stratum is large; we sample more heavily to compensate for the heterogeneity.
- Sampling in the stratum is inexpensive.

EXAMPLE 4.7 **Dollar stratification** is often used in accounting. The recorded book amounts are used to stratify the population. If you are auditing the loan amounts for a financial institution, stratum 1 might consist of all loans of more than \$1 million, stratum 2 might consist of loans between \$500,000 and \$999,999, and so on, down to the smallest stratum of loans less than \$10,000. Optimal allocation is often an efficient strategy for such a stratification: S_h^2 will be much larger in the strata with the large loan amounts, so optimal allocation will prescribe a higher sampling fraction for those strata. If the goal of the audit is to estimate the dollar discrepancy between the audited amounts and the amounts in the institution's books, an error in the recorded amount of one of the \$3 million loans is likely to contribute more to the audited difference than an error in the recorded amount of one of the \$3000 loans. In a survey such as this, you may even want to use sample size N_1 in stratum 1 so that each population unit in stratum 1 has probability 1 of appearing in the sample. ■

If all variances and costs are equal, proportional allocation is the same as optimal allocation. If we know the variances within each stratum and they differ, optimal allocation gives a smaller variance for the estimate of $\bar{y}_U$ than proportional allocation. But optimal allocation is a more complicated scheme; often the simplicity and self-weighting property of proportional allocation are worth the extra variance. In addition, the optimal allocation will differ for each variable being measured, whereas the proportional allocation depends only on the number of population units in each stratum.

Neyman allocation is a special case of optimal allocation, used when the costs in the strata (but not the variances) are approximately equal. Under Neyman allocation, n_h is proportional to $N_h S_h$. If the variances S_h^2 are specified correctly, Neyman allocation will give an estimator with smaller variance than proportional allocation.

EXAMPLE 4.8 The caribou survey in Example 4.2 used a form of optimal allocation to determine the n_h. Before taking the survey, the investigators obtained approximations of the caribou densities and distribution and then constructed strata to be relatively homogeneous in terms of population density. They set the total sample size as $n = 225$. They then used the estimated count in each stratum as a rough estimate of the standard deviation, with the result shown in Table 4.4. The first row contains the names of the spreadsheet columns, and the second row contains the formulas used to calculate the table. The investigators wanted the sampling fraction to be at least 1/3 in smaller strata, so they used the optimal allocation sample sizes in column E as a guideline for determining the sample sizes they actually used, in column F. ■

4.4.3 Allocation for Specified Precision Within Strata

Sometimes you are less interested in the precision of the estimate of the population total or mean for the whole population than in comparing means or totals among different strata. In that case, you would determine the sample size needed for the individual strata, using the guidelines in Section 2.5.

TABLE 4.4
Quantities Used for Designing the Caribou Survey in Example 4.8

	A	B	C	D	E	F
1	Stratum	N_h	s_h	$N_h s_h$	n_h	Sample size
2				B*C	D*225/\$D\$9	
3	A	400	3,000	1,200,000	96.26	98
4	B	30	2,000	60,000	4.81	10
5	C	61	9,000	549,000	44.04	37
6	D	18	2,000	36,000	2.89	6
7	E	70	12,000	840,000	67.38	39
8	F	120	1,000	120,000	9.63	21
9	total	699		2,805,000	225	211

EXAMPLE 4.9 The U.S. Postal Service often conducts surveys asking postal customers about their perceptions of the quality of mail service. The population of residential postal service customers is stratified by geographic area, and it is desired that the precision be ±3 percentage points, at a 95% confidence level, within each area. If there were no nonresponse, such a requirement would lead to sampling at least 1067 households in each stratum, as calculated in Example 2.9. Such an allocation is neither proportional, because the number of residential households varies a great deal from stratum to stratum, nor optimal in the sense of providing the greatest efficiency for estimating percentages for the whole population. It does, however, provide the desired precision within each stratum. ■

4.4.4 Determining Sample Sizes

The different methods of allocating observations to strata give the relative sample sizes n_h/n. After strata are constructed (see Section 4.5) and observations allocated to strata, Equation (4.3) can be used to determine the sample size necessary to achieve a prespecified variance. Because

$$V(\hat{t}_{\text{str}}) \le \frac{1}{n}\sum_{h=1}^{H} \frac{n}{n_h} N_h^2 S_h^2 = \frac{v}{n},$$

an approximate 95% CI if the fpc's can be ignored and if the normal approximation is valid will be $\hat{t}_{\text{str}} \pm z_{\alpha/2}\sqrt{v/n}$. Set $n = z_{\alpha/2}^2 v/e^2$ to achieve a desired confidence interval half-width e.

This approach requires knowledge of the values of S_h^2. An alternative approach, which works for any survey design, will be discussed in Section 7.5.

4.5 Defining Strata

One might wonder, since stratified sampling almost always gives higher precision than simple random sampling, why anyone would ever take SRSs. The answer is that stratification adds complexity to the survey, and the added complexity may not be worth a small gain in precision. In addition, to carry out a stratified sample, we need

more information: For each stratum, we need to know how many and which members of the population belong to that stratum. In general, we want stratification to be very efficient, or the strata to be subgroups we are interested in, before we will be willing to incur the additional administrative costs and complexity associated with stratifying.

Remember, stratification is most efficient when the stratum means differ widely; then the between sum of squares is large, and the variability within strata will be smaller. Consequently, when constructing strata we want the strata means to be as different as possible. Ideally, we would stratify by the values of y; if our survey is to estimate total business expenditures on advertising, we would like to put businesses that spent the most on advertising in stratum 1, businesses with the next highest level of advertising expenditures in stratum 2, and so on, until the last stratum contained businesses that spent nothing on advertising. The problem with this scheme is that we do not know the advertising expenditures for all the businesses while designing the survey—if we did, we would not need to do a survey at all! Instead, we try to find some variable closely related to y. For estimating total business expenditures on advertising, we might stratify by number of employees or size of the business and by the type of product or service. For farm income, we might use the size of the farm as a stratifying variable, since we expect that larger farms would have higher incomes.

Most surveys measure more than one variable, so any stratification variable should be related to many characteristics of interest. The U.S. Census Bureau's Current Population Survey, which measures characteristics relating to employment, stratifies the primary sampling units by geographic region, population density, racial composition, principal industry, and similar variables. In the Canadian Survey of Employment, Payrolls, and Hours, business establishments are stratified by industry, province, and estimated number of employees. The Nielsen television ratings stratify by geographic region, county size, and cable penetration, among other variables. If several stratification variables are available, use the variables associated with the most important responses.

The number of strata you choose depends on many factors—for example, the difficulty in constructing a sampling frame with stratifying information and the cost of stratifying. A general rule to keep in mind is: The less information, the fewer strata you should use. Thus, you should use an SRS when little prior information about the target population is available.

You can often collect preliminary data that can be used to stratify your design. If you are taking a survey to estimate the number of fish in a region, you can use physical features of the area that are related to fish density, such as depth, salinity, and water temperature. Or you can use survey information from previous years or data from a preliminary cruise to aid in constructing strata. In this situation, according to Saville, "Usually there will be no point in designing a sampling scheme with more than 2 or 3 strata, because our knowledge of the distribution of fish will be rather imprecise. Strata may be of different size, and each stratum may be composed of several distinct areas in different parts of the total survey area" (1977, 10). In a survey with more precise prior information, we will want to use more strata—many surveys are stratified to the point that only two sampling units are observed in each stratum.

For many surveys, stratification can increase precision dramatically and often well repays the effort used in constructing the strata. Example 4.10 describes how strata were constructed in one large-scale survey, the National Pesticide Survey.

EXAMPLE 4.10 Between 1988 and 1990, the U.S. Environmental Protection Agency (1990a, b) sampled drinking water wells to estimate the prevalence of pesticides and nitrate. When designing the National Pesticide Survey (NPS), the EPA scientists wanted a sample that was representative of drinking water wells in the United States. In particular, they wanted to guarantee that wells in the sample would have a wide range of levels of pesticide use and susceptibility to groundwater pollution. They also wanted to study two categories of wells: *community water systems* (CWSs), defined as "systems of piped drinking water with at least 15 connections and/or 25 or more permanent residents of the service area that have at least one working well used to obtain drinking water"; and *rural domestic wells,* "drinking water wells supplying occupied housing units located in rural areas of the United States, except for wells located on government reservations."

The following selections from the EPA describe how it chose the strata for the survey:

> In order to determine how many wells to visit for data collection, EPA first needed to identify approximately how many drinking water wells exist in the United States. This process was easier for community water systems than for rural domestic wells because a list of all public water systems, with their addresses, is contained in the Federal Reporting Data System (FRDS), which is maintained by EPA. From FRDS, EPA estimated that there were approximately 51,000 CWSs with wells in the United States. EPA did not have a comprehensive list of rural domestic wells to serve as the foundation for well selection, as it did for CWSs. Using data from the Census Bureau for 1980, EPA estimated that there were approximately 13 million rural domestic wells in the country, but the specific owners and addresses of these rural domestic wells were not known.
>
> EPA chose a survey design technique called "stratification" to ensure that survey data would meet its objectives. This technique was used to improve the precision of the estimates by selecting extra wells from areas with substantial agricultural activity and high susceptibility to ground-water pollution (vulnerability). EPA developed criteria for separating the population of CWS wells and rural domestic wells into four categories of pesticide use and three relative ground-water vulnerability measures. This design ensures that the range of variability that exists nationally with respect to the agricultural use of pesticides and ground-water vulnerability is reflected in the sample of wells.
>
> EPA identified five subgroups of wells for which it was interested in obtaining information. These subgroups were community water system wells in counties with relatively high average ground-water vulnerability; rural domestic wells in counties with relatively high average ground-water vulnerability; rural domestic wells in counties with high pesticide use; rural domestic wells in counties with both high pesticide use and relatively high average ground-water vulnerability; and rural domestic wells in "cropped and vulnerable" parts of counties (high pesticide use and relatively high ground-water vulnerability).
>
> Two of the most difficult design questions were determining how many wells to include in the Survey and determining the level of precision that would be sought for the NPS national estimates. These two questions were connected, because greater precision is usually obtained by collecting more data. Resolving these questions would have been simpler if the Survey designers had known in advance what proportion of

wells in the nation contained pesticides, but answering that question was one of the purposes of the Survey. Although many State studies have been conducted for specific pesticides, no reliable national estimates of well water contamination existed. EPA evaluated alternative precision requirements and costs for collecting data from different numbers of wells to determine the Survey size that would meet EPA's requirements and budget.

The Survey designers ultimately selected wells for data collection so that the Survey provided a 90 percent probability of detecting the presence of pesticides in the CWS wells sampled, assuming 0.5 percent of all community water system wells in the country contained pesticides. The rural domestic well Survey design was structured with different probabilities of detection for the several subgroups of interest, with the greatest emphasis placed on the cropped and vulnerable subcounty areas, where EPA was interested in obtaining very precise estimates of pesticide occurrence. EPA assumed that 1 percent of rural domestic wells in these areas would contain pesticides and designed the Survey to have about a 97 percent probability of detection in "cropped and vulnerable" areas if the assumption proved accurate. EPA concluded that sampling approximately 1,300 wells (564 public wells and 734 private wells) would meet the Survey's accuracy specifications and provide a representative national assessment of the number of wells containing pesticides.

Selecting Wells for the Survey. Because the exact number and location of rural domestic wells was unknown, EPA chose a survey design composed of several steps (stages) for those wells. The design began with a sampling of counties, and then characterized pesticide use and ground-water vulnerability for subcounty areas. This eventually allowed small enough geographic areas to be delineated to enable the sampling of individual rural domestic wells. This procedure was not needed for community water system wells, because their number and location were known.

The first step in well selection was common to both CWS wells and rural domestic wells. Each of the 3,137 counties or county equivalents in the U.S. was characterized according to pesticide use and ground-water vulnerability to ensure that the variability in agricultural pesticide use and ground-water vulnerability was reflected in the Survey. EPA used data on agricultural pesticide use obtained from a marketing research source and information on the proportion of the county area that was in agricultural production to rank agricultural pesticide use for each county as high, medium, low, or uncommon. Ground-water vulnerability of each county was estimated using a numerical classification system called Agricultural DRASTIC, which assesses seven factors: (depth of water, recharge, aquifer media, soil media, topography, impact of unsaturated zone, conductivity of the aquifer). The model was modified for the Survey to evaluate the vulnerability of aquifers to pesticide and nitrate contamination, and one of the subsidiary purposes of the Survey was to assess the effectiveness of the DRASTIC classification. Each area was evaluated and received a score of high, moderate, or low, based on information obtained from U.S. Geological Survey maps, U.S. Department of Agriculture soil survey maps and other resources from State agencies, associations, and universities. (1990a)

The procedure resulted in 12 strata for counties, as given in Table 4.5.

Stratification provides several advantages in this survey. It allows for more precise estimates of pesticide and nitrate concentrations in the United States as a whole, as

TABLE 4.5
Strata for National Pesticide Survey

Stratum	Pesticide Use	Groundwater Vulnerability (as Estimated by DRASTIC)	Number of Counties
1	High	High	106
2	High	Moderate	234
3	High	Low	129
4	Moderate	High	110
5	Moderate	Moderate	204
6	Moderate	Low	267
7	Low	High	193
8	Low	Moderate	375
9	Low	Low	404
10	Uncommon	High	186
11	Uncommon	Moderate	513
12	Uncommon	Low	416

SOURCE: Adapted from U.S. EPA 1990a, 3.

it is expected that the wells within a stratum are more homogeneous than the entire population of wells. Stratification ensures that wells for each level of pesticide use and groundwater vulnerability are included in the sample and allows estimation of pesticide concentration with a prespecified sample size in each stratum. The factorial design, with four levels of the factor *pesticide use* and three levels of the factor *groundwater vulnerability,* allows investigation of possible effects of each factor separately, and the interaction of the factors, on pesticide concentrations. ■

4.6 A Model for Stratified Sampling*

The one-way ANOVA model with fixed effects provides an underlying structure for stratified sampling. Here,

$$Y_{hj} = \mu_h + \varepsilon_{hj},$$

where the ε_{hj}'s are independent with mean 0 and variance σ_h^2. Then, as in Section 2.8, the least squares estimator of μ_h from units in the sample is the average of the sampled observations in stratum h.

Let the random variable

$$T_h = \sum_{j=1}^{N_h} Y_{hj}$$

represent the total in stratum h and the random variable

$$T = \sum_{h=1}^{H} T_h$$

represent the overall total.

From Section 2.8, the best linear unbiased estimator for T_h is

$$\hat{T}_h = \frac{N_h}{n_h} \sum_{j \in \mathcal{S}_h} Y_{hj}.$$

Then, from the results shown for simple random sampling in Section 2.8,

$$E_M[\hat{T}_h - T_h] = 0$$

and

$$E_M[(\hat{T}_h - T_h)^2] = N_h^2 \frac{\sigma_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right).$$

Since we sample independently in the strata,

$$\begin{aligned} E_M[(\hat{T} - T)^2] &= E_M\left[\left\{\sum_{h=1}^{H} (\hat{T}_h - T_h)\right\}^2\right] \\ &= E_M\left[\sum_{h=1}^{H} (\hat{T}_h - T_h)^2 + \sum_{h=1}^{H} \sum_{k \neq h} (\hat{T}_h - T_h)(\hat{T}_k - T_k)\right] \\ &= E_M\left[\sum_{h=1}^{H} (\hat{T}_h - T_h)^2\right] \\ &= \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{\sigma_h^2}{n_h}. \end{aligned}$$

The theoretical variance σ_h^2 can be estimated by s_h^2. Adopting this model results in the same estimates for t and its standard error as found under randomization theory in Equations (4.1) and (4.4). If a different model is used, however, then different estimates are obtained.

4.7 Poststratification

Suppose a sampling frame lists all households in an area, and you would like to estimate the average amount spent on food in a month. One desirable stratification variable might be household size because large households might be expected to have higher food bills than smaller households. From U.S. census data, the distribution of household size in the region is known:

Number of Persons in Household	Percentage of Households
1	25.75
2	31.17
3	17.50
4	15.58
5+	10.00

The sampling frame, however, does not include information on household size—it only lists the households.

Without additional information, you cannot design an intelligent stratified sampling plan. You can, however, take an SRS and record the amount spent on food as well as the household size for each household in your sample. If n, the size of the SRS, is large enough, then the sample is likely to resemble a stratified sample with proportional allocation: We would expect about 26% of the sample to be one-person households, about 31% to be two-person households, and so on.

Considering the different household-size groups to be different domains, we can use the methods from Section 3.3 to estimate the average amount spent on groceries for each domain: Take an SRS of size n. Let $n_1, n_2, \ldots, n_H$ be the numbers of units sampled in the various household-size groups (domains) and $\bar{y}_1, \ldots, \bar{y}_H$ be the sample means for the groups.

After the observations are taken, form a "stratified" estimate of $\bar{y}_U$:

$$\bar{y}_{\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h. \tag{4.14}$$

If (1) N_h/N is known, (2) n_h is reasonably large ($\geq$30 or so), and (3) n is large, then we can use the variance for proportional allocation as a good approximation to the variance:

$$\hat{V}(\bar{y}_{\text{post}}) \approx \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} \frac{N_h}{N} \frac{s_h^2}{n}. \tag{4.15}$$

WARNING: Poststratification can be *dangerous* if you indulge in data snooping: You can obtain arbitrarily small variances if you choose the strata after seeing the data, just as you can always obtain statistical significance if you decide on your null and alternative hypotheses after looking at the data. Poststratification is most often used to correct for the effects of differential nonresponse in the poststrata (see Chapter 8).

4.8 Quota Sampling

Many samples that masquerade as stratified random samples are actually quota samples. In **quota sampling,** the population is divided into different subpopulations just as in stratified random sampling, but with one important difference: Probability sampling is not used to choose individuals in the subpopulation for the sample. In extreme versions of quota sampling, choice of units in the sample is entirely at the discretion of the interviewer, so a sample of convenience is chosen within each subpopulation.

In quota sampling, specified numbers (quotas) of particular types of population units are required in the final sample. For example, to obtain a quota sample with $n = 3000$, you might specify that the sample contain 1000 white males, 1000 white females, 500 men of color, and 500 women of color, but you might give no further

instructions about how these quotas are to be filled. Thus, quota sampling is not a form of probability sampling—we do not know the probabilities with which each individual is included in the sample. It is often used when probability sampling is impractical, overly costly, or considered unnecessary, or when the persons designing the sample just do not know any better.

The big drawback of quota sampling is that we do not know if the units chosen for the sample exhibit selection bias. If selection of units is totally up to the interviewer, she or he is likely to choose the most accessible members of the population—for instance, persons who are easily reached by telephone, households without menacing dogs, or areas of the forest close to the road. The most accessible members of a population are likely to differ in a systematic way from less accessible members. Thus, unlike in stratified random sampling, we cannot say that the estimate of the population total from quota sampling is unbiased over repeated sampling—one of our usual criteria of goodness in probability samples. In fact, in quota samples we cannot measure sampling error over repeated samples and we have no way of estimating the bias from the sample data. Since selection of units is up to the individual interviewer, we cannot expect that repeating the sample will give similar results. Thus, anyone drawing inferences from a quota sample must necessarily take a model-based approach.

EXAMPLE 4.11 The 1945 survey on reading habits taken for the Book Manufacturer's Institute (Link and Hopf 1946), like many surveys in the 1940s and 1950s, used a quota sample. Some of the classifications used to define the quota classes were area, city size, age, sex, and socioeconomic status; a local supervising psychologist in each city determined the blocks of the city in which interviewers were to interview people from a specified socioeconomic group. The interviewers were then allowed to choose the specific households to be interviewed in the designated city blocks.

The quota procedure followed in the survey did not result in a sample that reflected demographic characteristics of the 1945 U.S. population. The following table compares the educational background of the survey respondents with figures from the 1940 U.S. census, adjusted to reflect the wartime changes in population:

Distribution by Educational Levels	4000 People Interviewed (%)	U.S. Census, Urban and Rural Nonfarm (%)
8th grade or less	28	48
1–3 years high school	18	19
4 years high school	25	21
1–3 years college	15	7
4 or more years college	13	5

SOURCE: Link and Hopf 1946.

The oversampling of better-educated persons casts doubt on many of the statistics given in the book. The study concluded that 31% of "active readers" (those who had read at least one book in the past month) had bought the last book they read and that

25% of all last books read by active readers cost $1 or less. Who knows whether a stratified random sample would have given the same results? ■

In the 1948 U.S. presidential election, all major polls printed just a few days before the election predicted that Dewey would handily defeat Truman. In fact, of course, Truman won the election. According to Mosteller et al. (1949), one problem of those polls was that they all used quota sampling, not a probability-based method. This 1948 polling debacle spurred many American survey organizations to turn away from quota sampling, at least for a few years.

Many electoral polls in Britain used probability samples in the 1960s and 1970s. Probability samples, however, were much more expensive than quota samples, and quota samples used in the 1970s gave accurate predictions of the election results, so several large polling organizations went back to quota sampling (Taylor 1995). The polls that erred in predicting the winner in the 1992 British general election all used quota methods in selecting persons to interview in their homes or in the street; the primary quota classes used were sex, age, socioeconomic class, and employment status. Although we may never know exactly what went wrong in those polls (see Crewe 1992 for some other explanations), the use of quota samples may have played a part—if interviewing persons "in the street," it is certainly plausible that persons from a quota class who are accessible differ from persons who are less accessible.

Although quota sampling is not as good as probability sampling under ideal conditions, it will usually give much better results than the convenience samples that are often taken because it at least forces the inclusion of members of the different quota groups. Quota samples have the advantage of being less expensive than probability samples. The quality of the data from quota samples can be improved by allowing the interviewer less discretion in the choice of persons or households to be included in the sample. Many survey organizations use probability sampling along with quotas; they use probability sampling to select small blocks of potential respondents and then take a quota sample within each block, using variables such as age, sex, and race to define the quota classes.

A quota sample performs unfavorably compared with a stratified random sample when there is no nonresponse in the stratified random sample. When there is nonresponse, the comparison is unclear. Quota sampling can be considered as a substitution method for dealing with nonresponse, as is considered in Chapter 8: A nonrespondent is replaced by another person in the same quota class.

Because we do not know the probabilities with which units were sampled, we must take a model-based approach when analyzing data from a quota sample. The model generally adopted is that of Section 4.6—within each subclass the random variables generating the subpopulation are independent and identically distributed. Such a model implies that any selection of units from the subclass will give a representative sample; if the model holds, then quota sampling will likely give good estimates of the population quantity. If the model does not hold, then the estimates from quota sampling may be badly biased.

Deville (1991, 177) argues that quota samples may be useful for market research, when the organization requesting the survey is aware of the model being used. Persons collecting official statistics about crime, unemployment, or other matters that may be debated should use probability samples, however.

EXAMPLE 4.12 Sanzo et al. (1993) used a combination of stratified random sampling and quota sampling for estimating the prevalence of *Coxiella burnetii* infection within the Basque country in northern Spain. *Coxiella burnetii* can cause Q fever, which can lead to complications such as heart and nerve damage. Reviews of Q fever patient records from Basque hospitals showed that about three-fourths of the victims were male, about one-half were between 16 and 30 years old, and victims were disproportionately likely to be from areas with low population density.

The authors stratified the target population by population density and then randomly selected health-care centers from the three strata. In selecting persons for blood testing, however, "a probabilistic approach was rejected as we considered that the refusal rate of blood testing would be high" (p. 1185). Instead, they used quota sampling to balance the sample by age and gender; physicians asked patients who needed laboratory tests whether they would participate in the study and recruited subjects for the study until the desired sample sizes in the six quota groups were reached for each stratum.

Because a quota sample was taken instead of a probability sample, persons analyzing the data must make strong assumptions about the representativeness of the sample in order to apply the results to the general population of the Basque country. First, the assumption must be made that persons attending a health clinic for laboratory tests (the sampled population of the study) are neither more nor less likely to be infected than persons who would not be visiting the clinic. Second, one must assume that persons who are requested and agree to do the study are similar in terms of the infection to persons in the same quota class having laboratory tests that do not participate in the study. These are strong assumptions. The authors of the article argue that the assumptions are justified, but of course they cannot prove that the assumptions hold unless follow-up investigations are done.

If they had taken a probability sample of persons instead of the quota sample, they would not have had to make these strong assumptions. A probability sample of persons, however, would have been exhorbitantly expensive when compared with the quota design used, and a probability sample would also have taken longer to design and implement. With the quota sample, the authors could collect information about the public health problem; it is unclear whether the results can be generalized to the entire population, but the data do provide a great deal of quick information on the prevalence of infection that can be used in future investigation of who is likely to be infected, and why. ■

4.9 Exercises

1 What stratification variable(s) would you use for each of the following situations?

a A political poll to estimate the percentage of registered voters in Arizona who approve of the job the governor is doing

b A telephone survey of students at your university, to estimate the total amount of money students spend on textbooks

c A sample of high schools in New York City, to estimate what percentage of high schools offer one or more classes in computer programming

d A sample of public libraries in California, to study the availability of computer resources and the per capita expenditures

e A survey of anglers visiting a freshwater lake, to learn about which species of fish are preferred

f An aerial survey to estimate the number of walrus in the pack ice near Alaska between 173° east and 154° west longitude

g A sample of prime-time (7–10 P.M., Monday through Saturday; 6–10 P.M., Sunday) TV programs on CBS, to estimate the average number of promotional announcements (ads for other programming on the station) per hour of broadcast

2 The data set agstrat.dat also contains information on other variables. For each of the following quantities, plot the data and estimate the population mean for that variable along with its standard error. Give a 95% CI for each estimate. Compare your answers with those from the SRS in Exercise 11 in Chapter 2.

a Number of acres devoted to farms, 1987

b Number of farms, 1992

c Number of farms with 1000 acres or more, 1992

d Number of farms with 9 acres or fewer, 1992

3 Hard-shell clams can be sampled by using a dredge. Clams do not tend to be uniformly distributed in a body of water, however, because some areas provide better habitat than others. Thus, taking an SRS is likely to result in a large estimated variance for the number of clams in an area. Russell (1972) used stratified random sampling to estimate the total number of bushels of hard-shell clams *(Mercenaria mercenaria)* in Narragansett Bay, Rhode Island. The area of interest was divided into four strata based on preliminary surveys that identified areas in which clams were abundant. Then, n_h dredge tows were made in stratum h, for $h = 1, 2, 3, 4$. The acreage for each stratum was known, and Russell calculated that the area fished during a standard dredge tow was 0.039 acre—that is, 25.6 dredge tows would fish 1 acre.

a Here are the results from the survey taken before the commercial season. Estimate the total number of bushels of clams in the area and give the standard error of your estimate. HINT: First calculate N_h, the number of dredge tows needed to cover stratum h.

Stratum	Area (Acres)	Number of Tows Made	Average Number of Bushels per Tow	Sample Variance for Stratum
1	222.81	4	0.44	0.068
2	49.61	6	1.17	0.042
3	50.25	3	3.92	2.146
4	197.81	5	1.80	0.794

b Another survey was performed at the end of the commercial season. In this survey, strata 1, 2, and 3 were collapsed into a single stratum, called stratum 1 below. Estimate the total number of bushels of clams (with standard error) at the end of the season.

Stratum	Area (Acres)	Number of Tows Made	Average Number of Bushels per Tow	Sample Variance for Stratum
1	322.67	8	0.63	0.083
4	197.81	5	0.40	0.046

4 Return to the hypothetical population in Example 3.4. Now, instead of using x as an auxiliary variable in ratio estimation, use it as a stratification variable: A population unit is in stratum 1 if $x \leq 5$ and in stratum 2 if $x > 5$. With this stratification, $N_1 = N_2 = 4$. The population is as follows:

Unit Number	Stratum	y
1	1	1
2	1	2
3	1	4
8	1	8
4	2	4
5	2	7
6	2	7
7	2	7

Consider the stratified sampling design in which $n_1 = n_2 = 2$.

a Write out all possible SRSs of size 2 from stratum 1 and find the probability of each sample. Do the same for stratum 2.

b Using your work in part (a), find the sampling distribution of $\hat{t}_{\text{str}}$.

c Find the mean and variance of the sampling distribution of $\hat{t}_{\text{str}}$. How do these compare with the mean and variance in Examples 2.1 and 3.4?

5 Suppose a city has 90,000 dwelling units, of which 35,000 are houses, 45,000 are apartments, and 10,000 are condominiums. You believe that the mean electricity usage is about twice as much for houses as for apartments or condominiums and that the standard deviation is proportional to the mean.

a How would you allocate a sample of 900 observations if you want to estimate the mean electricity consumption for all households in the city?

b Now suppose that you want to estimate the overall proportion of households in which energy conservation is practiced. You have strong reason to believe that about 45% of house dwellers use some sort of energy conservation and that the corresponding percentages are 25% for apartment dwellers and 3% for condominium residents. What gain would proportional allocation offer over simple random sampling?

c Someone else has taken a small survey, using an SRS, of energy usage in houses. On the basis of the survey, each house is categorized as having electric heating or some other kind of heating. The January electricity consumption in kilowatt-hours for each house is recorded (y_i) and the results are given below:

Type of Heating	Number of Houses	Sample Mean	Sample Variance
Electric	24	972	202,396
Nonelectric	36	463	96,721
Total	60		

From other records, it is known that 16,450 of the 35,000 houses have electric heating, and 18,550 have nonelectric heating.

i Using the sample, give an estimate and its standard error of the proportion of houses with electric heating. Does your 95% CI include the true proportion?

ii Give an estimate and its standard error of the average number of kilowatt-hours used by houses in the city. What type of estimator did you use, and why did you choose that estimator?

6 A public opinion researcher has a budget of \$20,000 for taking a survey. She knows that 90% of all households have telephones. Telephone interviews cost \$10 per household; in-person interviews cost \$30 each if all interviews are conducted in person and \$40 each if only nonphone households are interviewed in person (because there will be extra travel costs). Assume that the variances in the phone and nonphone strata are similar and that the fixed costs are $c_0 = \$5000$. How many households should be interviewed in each stratum if

a All households are interviewed in person.

b Households with a phone are contacted by telephone and households without a phone are contacted in person.

7 For Example 4.3, construct a data set with 3835 observations. Include three columns: column 1 is the stratum number (from 1 to 7), column 2 contains the response variable of gender (0 for males and 1 for females), and column 3 contains the sampling weight N_h/n_h for each observation. Using columns 2 and 3 along with (4.10), calculate $\hat{p}_{\text{str}}$. Is it possible to calculate SE($\hat{p}_{\text{str}}$) by using only columns 2 and 3, with no additional information?

8 The survey in Example 4.3 collected much other data on the subjects. Another of the survey's questions asked whether the respondent agreed with the following statement: "When I look at a new issue of my discipline's major journal, I rarely find an article

that interests me." The results are as follows:

Discipline	Agree (%)
Literature	37
Classics	23
Philosophy	23
History	29
Linguistics	19
Political science	43
Sociology	41

a What is the sampled population in this survey?

b Find an estimate of the proportion of persons in the sampled population that agree with the statement and give the standard error of your estimate.

9 Construct a small population and stratification for which $V(\hat{t}_{str})$ using proportional allocation is larger than the variance that would be obtained by taking an SRS with the same number of observations. HINT: Use (4.11).

10 In Exercise 8 of Chapter 2, data on numbers of publications were given for an SRS of 50 faculty members. Not all departments, however, were represented in the SRS. The SRS contained several faculty members from psychology and from chemistry but none from foreign languages. The following data are from a stratified sample of faculty, using the areas biological sciences, physical sciences, social sciences, and humanities as the strata. Proportional allocation was used in this sample.

Stratum	Number of Faculty Members in Stratum	Number of Faculty Members in Sample
Biological sciences	102	7
Physical sciences	310	19
Social sciences	217	13
Humanities	178	11
Total	807	50

The frequency table for number of publications in the strata is given below.

Number of Refereed Publications	Number of Faculty Members: Biological	Physical	Social	Humanities
0	1	10	9	8
1	2	2	0	2
2	0	0	1	0
3	1	1	0	1
4	0	2	2	0
5	2	1	0	0
6	0	1	1	0
7	1	0	0	0
8	0	2	0	0

a Estimate the total number of refereed publications by faculty members in the college and give the standard error.

b How does your result from part (a) compare with the result from the SRS in Exercise 8 of Chapter 2?

c Estimate the proportion of faculty with no refereed publications and give the standard error.

d Did stratification increase precision in this example? Explain why you think it did or did not.

11 Lydersen and Ryg (1991) used stratification techniques to estimate ringed seal populations in Svalbard fjords. The 200-km^2 study area was divided into three zones: Zone 1, outer Sassenfjorden, was covered with relatively new ice during the study period in March 1990 and had little snow cover; zone 3, Tempelfjorden, had a stable ice cover throughout the year; zone 2, inner Sassenfjorden, was intermediate between the stable zone 3 and the unstable zone 1. Ringed seals need good ice to establish territories with breeding holes, and snow cover enables females to dig out birth lairs. Thus, it was thought that the three zones would have different seal densities.

To select the sample, investigators divided the entire region into 200 1-km^2 areas; "a sampling grid covering 20% of the total area was made . . . by picking 40 numbers between one and 200 with the random number generator." In each sampled area, Imjak the Siberian husky tracked seal structures by scent; the number of breathing holes in each sampled square was recorded. A total of 199 breathing holes were located in zones 1–3. The data (reconstructed from information given in the paper) are in the file seals.dat.

The following table gives the number of plots, and the number of plots sampled, in each zone:

Zone	Number of Plots	Plots Sampled
1	68	17
2	84	12
3	48	11
Total	200	40

a Is this a stratified random sample, or a poststratified SRS? Explain.

b Estimate the total number of breathing holes in the study region, along with its standard error.

c If you were designing this survey, how would you allocate observations to strata if the goal was to estimate the total number of breathing holes? If the goal was to compare the density of breathing holes in the three zones?

12 Proportional allocation was used in the stratified sample in Example 4.1. It was noted, however, that variability was much higher in the West than in the other regions. Using the estimated variances in Example 4.1 and assuming that the sampling costs are the same in each stratum, find an optimal allocation for a stratified sample of size 300.

13 Select a stratified random sample of size 300 from the data in the file agpop.dat, using your allocation in Exercise 12. Estimate the total number of acres devoted to

farming in the United States and give the standard error of your estimate. How does this standard error compare with that found in Example 4.1?

14 Burnard (1992) sent a questionnaire to a stratified sample of nursing tutors and students in Wales, to study what the tutors and students understood by the term *experiential learning*. The population size and sample size obtained for each of the four strata are given below:

Stratum	Population Size	Sample Size
General nursing tutors (GT)	150	109
Psychiatric nursing tutors (PT)	34	26
General nursing students (GS)	2680	222
Psychiatric nursing students (PS)	570	40
Total	3434	397

Respondents were asked which of the following techniques could be identified as experiential learning methods; the number of students and tutors in each group who identified the method as an experiential learning method are given below:

Method	GS	PS	PT	GT
Role play	213	38	26	104
Problem-solving activities	182	33	22	95
Simulations	95	20	22	64
Empathy-building exercises	89	25	20	54
Gestalt exercises	24	4	5	12

Estimate the overall percentage of nursing students and tutors who identify each of these techniques as *experiential learning*. Be sure to give standard errors for your estimates.

15 Kruuk et al. (1989) used a stratified sample to estimate the number of otter *(Lutra lutra)* dens along the 1400-km coastline of Shetland, UK. The coastline was divided into 242 (237 that were not predominantly buildings) 5-km sections, and each section was assigned to the stratum whose terrain type predominated. Sections were then chosen randomly from the sections in each stratum. In each section chosen, investigators counted the total number of dens in a 110-m-wide strip along the coast.

The data are in the file otters.dat. The population sizes for the strata are as follows:

Stratum	Total Sections	Sections Counted
1 Cliffs over 10 m	89	19
2 Agriculture	61	20
3 Not 1 or 2, peat	40	22
4 Not 1 or 2, nonpeat	47	21

a Estimate the total number of otter dens along the coast in Shetland, along with a standard error for your estimate.

b Discuss possible sources of bias in this study. Do you think it is possible to avoid all selection and measurement bias?

16 Marriage and divorce statistics are compiled by the National Center for Health Statistics and published in volumes of *Vital Statistics of the United States*. State and local officials provide NCHS with annual counts of marriages and divorces in each county. In addition, some states send computer tapes of additional data or microfilm copies of marriage or divorce certificates. These additional data are used to calculate statistics about age at marriage or divorce, previous marital status of marrying couples, and children involved in divorce. In 1987, if a state sent a computer tape, all records were included in the divorce statistics; if a state sent microfilm copies, a specified fraction of the divorce certificates was randomly sampled and data recorded. The sampling rates (probabilities of selection) and number of records sampled in each state in the divorce registration area for 1987 are in the file divorce.dat.

a How many divorces were there in the divorce registration area in 1987? HINT: Use the sampling weights.

b Why did NCHS use different sampling rates in different states?

c Estimate the total number of divorces granted to men aged 24 or less; to women aged 24 or less. Give 95% CIs for your estimates.

d In what proportion of all divorces is the husband between 40 and 49 years old? In what proportion is the wife between 40 and 49 years old? Give confidence intervals for your estimates.

17 Jackson et al. (1987) compared the precision of systematic and stratified sampling for estimating the average concentration of lead and copper in the soil. The 1-km^2 area was divided into 100-m squares, and a soil sample was collected at each of the resulting 121 grid intersections. Summary statistics from this systematic sample are given below:

Element	n	Average (mg kg^{-1})	Range (mg kg^{-1})	Standard Deviation (mg kg^{-1})
Lead	121	127	22–942	146
Copper	121	35	15–90	16

The investigators also poststratified the same region. Stratum A consisted of farmland away from roads, villages, and woodlands. Stratum B contained areas within 50 m of roads and was expected to have larger concentrations of lead. Stratum C contained the woodlands, which were also expected to have larger concentrations of lead because the foliage would capture airborne particles. The data on concentration of lead and copper were not used in determining the strata. The data from the grid points falling

in each stratum are in the following table:

Element	Stratum	n_h	Average (mg kg^{-1})	Range (mg kg^{-1})	Standard Deviation (mg kg^{-1})
Lead	A	82	71	22–201	28
Lead	B	31	259	36–942	232
Lead	C	8	189	88–308	79
Copper	A	82	28	15–68	9
Copper	B	31	50	22–90	18
Copper	C	8	45	31–69	15

a Calculate a 95% CI for the average concentration of lead in the area, using the systematic sample. (You may assume that this sample behaves like an SRS.) Repeat for the average concentration of copper.

b Now use the poststratified sample and find 95% CIs for the average concentration of lead and copper. How do these compare with the confidence intervals in part (a)? Do you think that using stratification in future surveys would increase precision?

18 In Exercise 17 the sample size in each stratum was proportional to the area of the stratum. Using the sample standard deviations, what would an optimal allocation be for taking a stratified random sample with 121 observations? Is the optimal allocation the same for copper and lead?

19 Wilk et al. (1977) report data on the number and types of fish and environmental data for the area of the Atlantic continental shelf between eastern Long Island, New York, and Cape May, New Jersey. The ocean survey area was divided into strata based on depth. Sampling was done at a higher rate close to shore than farther away from shore: "In-shore strata (0–28 m) were sampled at a rate of approximately one station per 515 km^2 and off-shore strata (29–366 m) were sampled at a rate of approximately one station per 1,030 km^2" (p. 1). Thus, each record in strata 3–6 represents twice as much area as each record in strata 1 and 2. In calculating average numbers of fish caught and numbers of species, we can use a relative sampling weight of 1 for strata 1 and 2, and weight 2 for strata 3–6.

Stratum	Depth (m)	Relative Sampling Weight
1	0–19	1
2	20–28	1
3	29–55	2
4	56–100	2
5	111–183	2
6	184–366	2

The file nybight.dat contains data on the total catch for sampling stations visited in June 1974 and June 1975.

a Construct side-by-side boxplots of the number of fish caught in the trawls in June 1974. Does there appear to be a large variation among the strata?

b Calculate estimates of the average number and average weight of fish caught per haul in June 1974, along with the standard error.

c Calculate estimates of the average number and average weight of fish caught per haul in June 1975, along with the standard error.

d Is there evidence that the average weight of fish caught per haul differs between June 1974 and June 1975? Answer using an appropriate hypothesis test.

20 In January 1995 the Office of University Evaluation at Arizona State University surveyed faculty and staff members to find out their reaction to the closure of the university during the winter break in 1994. Faculty and staff in academic units that were closed during the winter break were divided into four strata and subsampled:

Stratum Number	Employee Type	Population Size (N_h)	Sample Size
1	Faculty	1374	500
2	Classified staff	1960	653
3	Administrative staff	252	98
4	Academic professional	95	95

Questionnaires were sent through campus mail to persons in strata 1–4; the sample size in the above table is the number of questionnaires mailed in each stratum. We'll come back to the issue of nonresponse in this survey in Chapter 8; for now, just analyze the respondents in the stratified sample of employees in closed units; the data for the 985 survey respondents are found in the file winter.dat. For this exercise, look at the answers to the question "Would you want to have Winter Break closure again?" (variable *breakaga*).

a Not all persons in the survey responded to the question. Find the number of persons who responded to the question in each of the four strata. For this exercise, use these values as the n_h.

b Use (4.6) and (4.7) to estimate the proportion of faculty and staff that would answer yes to the question "Would you want to have Winter Break closure again?" and give the standard error.

c Create a new variable, in which persons who respond yes to the question take on the value 1, persons who respond no to the question take on the value 0, and persons who do not respond are either left blank (if you are using a spreadsheet) or assigned the missing value code (if you are using statistical software). Construct a column of sampling weights N_h/n_h for the observations in the sample. (The sampling weight will be zero or missing for nonrespondents.) Now use (4.10) to estimate the proportion of faculty and staff that would answer yes to the question "Would you want to have Winter Break closure again?"

d Using the column of 0s and 1s you constructed in the previous question, find s_h^2 for each stratum by calculating the sample variance of the observations in that stratum. Now use (4.5) to calculate the standard error of your estimate of the proportion. Why is your answer the same as you calculated in part (b)?

e Stratification is sometimes used as a method of dealing with nonresponse. Calculate the response rates (the number of persons who responded divided by the number of questionnaires mailed) for each stratum. Which stratum has the lowest response rate for this question? How does stratification treat the nonrespondents?

21 A stratified sample is being designed to estimate the prevalence p of a rare characteristic—say, the proportion of residents in Milwaukee who have Lyme disease. Stratum 1, with N_1 units, has a high prevalence of the characteristic; stratum 2, with N_2 units, has low prevalence. Assume that the cost to sample a unit (for example, the cost to select a person for the sample and determine whether he or she has Lyme disease) is the same for each stratum and that at most 2000 units are to be sampled.

a Let p_1 and p_2 be the respective proportions in stratum 1 and stratum 2 with the rare characteristic. If $p_1 = 0.10$, $p_2 = 0.03$, and $N_1/N = 0.4$, what are n_1 and n_2 under optimal allocation?

b If $p_1 = 0.10$, $p_2 = 0.03$, and $N_1/N = 0.4$, what is $V(\hat{p}_{\text{str}})$ under proportional allocation? Under optimal allocation? What is the variance if you take an SRS of 2000 units from the population?

c (Use a spreadsheet for this part of the exercise.) Now fix $p = 0.05$. Let p_1 range from 0.05 to 0.50, and N_1/N range from 0.01 to 0.50 (these two values then determine the value of p_2). For each combination of p_1 and N_1/N, find the optimal allocation and the variance under both proportional allocation and optimal allocation. Also find the variance from an SRS of 2000 units. When does the optimal allocation give a substantial increase in precision when compared to proportional allocation? When compared to an SRS?

***22** (Requires calculus.) Show that the variance of $\hat{t}_{\text{str}}$ is minimized for a fixed cost with the cost function in (4.12) when $n_h \propto N_h S_h/\sqrt{c_h}$, as in (4.13). HINT: Use Lagrange multipliers.

23 Suppose the Arizona Department of Health wishes to take a survey of 2-year-olds whose families receive medical assistance, to determine the proportion who have been immunized. The medical care is provided by several different health-care organizations, and the state has 15 counties. Table 4.6 shows the population number of 2-year-olds for each county/organization combination. The sample is to be stratified by county and organization. It is desired to select sample sizes for each combination so that

a The margin of error for estimating percentage immunized is 0.05 or less when the data are tabulated for each county (summing over all health-care organizations).

b The margin of error for estimating percentage immunized is 0.05 or less when the data are tabulated for each health-care organization (summing over all counties).

c At least two children (fewer, of course, if the cell does not have two children) are selected from every cell.

Note that for this problem, as for many survey designs, many different designs would be possible.

TABLE 4.6
Table for Exercise 23

	A	B	C	D	E	Other	Total
Apache	1	13	19	0	0	94	127
Cochise	2	5	0	637	40	0	694
Coconino	1	6	0	125	0	289	421
Gila	0	2	51	151	0	0	204
Graham	0	2	0	63	0	143	208
Greenlee	0	0	0	58	0	0	58
Maricopa	118	169	0	3,732	2,675	5,105	11,799
Mohave	4	6	0	44	0	476	530
Navajo	2	5	132	124	0	0	263
Pima	62	26	0	1,097	727	1,786	3,698
Pinal	5	10	13	22	360	478	888
Santa Cruz	0	5	0	118	150	0	273
Yavapai	7	8	0	173	0	198	386
Yuma	5	5	0	837	0	0	847
LaPaz	0	1	0	89	0	0	90
Total	217	263	215	7,270	3,952	8,569	20,486

SURVEY Exercises

24 In the quest to estimate the average price a household in Stephens County is willing to pay for cable TV service, we are fortunate to know a great deal about some demographic aspects of the county, as given in the district map and tables in Appendix A. According to the SURVEY assumptions, what information might be used to stratify Stephens County in order to improve the precision of estimates? Are any other reasons for stratification relevant to Stephens County?

25 Use any considerations you like to divide Stephens County into strata. Your stratification should divide Lockhart City into approximately five strata. Why did you choose your stratification variable? Count the total number of households in each of your strata. (You may use the ADDGEN program to do this.)

The remainder of these exercises concern Lockhart City *only*.

26 Using ADDGEN, generate a stratified random sample of size 200 from Lockhart City with your stratification in Exercise 25 and proportional allocation. Find the responses using the SURVEY program. Estimate the average price a household in Lockhart City is willing to pay for cable service and the average number of TVs per household in Lockhart City. How do these estimates compare with those obtained with simple random sampling and sample mean and ratio estimates? Which estimates are the most precise?

27 Pilot studies are often used to estimate S_h. In this case we are fortunate to have a very large pilot study from the sample of size 200 used in Exercise 28 in Chapter 2.

Divide your sample from Chapter 2 into the strata you chose above and thus obtain estimates of the variances S_h^2 in each of the strata for the average price a household is willing to pay for cable TV service.

28 The sampling costs for Stephens County are given in Appendix A. Using your estimates of S_h, optimally allocate a sample of size 200 to estimate the average price a household in Lockhart City is willing to pay for cable TV service. Using that allocation, take a stratified random sample of Lockhart City and estimate the average price a household is willing to pay for cable TV service and the average number of TVs per household.

29 Under what conditions can optimal allocation be expected to perform much better than proportional allocation? Do these conditions occur in Lockhart City? Comment on the relative performance that you observed between these two allocations.

30 Using the variances estimated in Exercise 28 of Chapter 2, what sample size would be needed with simple random sampling to achieve the same precision in estimating the average price a household is willing to pay as a stratified sample of size 200 using the strata you have designed and optimal allocation? Proportional allocation?

31 Are there any deficiencies in your design? How would you correct them if you were to do this exercise a second time?

5

Cluster Sampling with Equal Probabilities

"But averages aren't real," objected Milo; "they're just imaginary."

"That may be so," he agreed, "but they're also very useful at times. For instance, if you didn't have any money at all, but you happened to be with four other people who had ten dollars apiece, then you'd each have an average of eight dollars. Isn't that right?"

"I guess so," said Milo weakly.

"Well, think how much better off you'd be, just because of averages," he explained convincingly. "And think of the poor farmer when it doesn't rain all year: if there wasn't an average yearly rainfall of 37 inches in this part of the country, all his crops would wither and die."

It all sounded terribly confusing to Milo, for he had always had trouble in school with just this subject.

"There are still other advantages," continued the child. "For instance, if one rat were cornered by nine cats, then, on the average, each cat would be 10 per cent rat and the rat would be 90 per cent cat. If you happened to be a rat, you can see how much nicer it would make things."

—Norton Juster, *The Phantom Tollbooth*

In all the sampling procedures discussed so far, we have assumed that the population is given and all we must do is reach in and take a suitable sample of units. But units are not necessarily nicely defined, even when the population is. There may be several ways of listing the units, and the unit size we choose may very well contain smaller subunits.

Suppose we want to find out how many bicycles are owned by residents in a community of 10,000 households. We could take a simple random sample (SRS) of 400 households, or we could divide the community into blocks of about 20 households each and sample every household (or subsample some of the households) in each of 20 blocks selected at random from the 500 blocks in the community. The latter plan is an example of cluster sampling. The blocks are the **primary sampling units** (psu's), or **clusters.** The households are the **secondary sampling units** (ssu's); often the ssu's are the elements in the population.

The cluster sample of 400 households is likely to give less precision than an SRS of 400 households; some blocks of the community are composed mainly of families (with more bicycles), whereas the residents of other blocks are mainly retirees (with