Survey Sampling
1. Introduction

1.1 A sample controversy

Shere Hites book Women and Love: A Cultural Revolution in Progress (1987) had a number of widely quoted results:

- 84% of women are not satisfied emotionally with their relationships (p. 804).

- 70% of all women married five or more years are having sex outside of their marriages (p. 856).

- 95% of women report forms of emotional and psychological harassment from men with whom they are in love relationships (p. 810).

- 84% of women report forms of condescension from the men in their love relationships (p. 809).
Hite erred in generalizing these results to all women, whether they participated in the survey or not, and in claiming that the percentages above applied to all women.

- The sample was self-selected. Hite mailed 100,000 questionnaires; of these, 4.5% were returned.

- The questionnaires were mailed to organizations, whose members have joined “all-women” group.

- The survey has 127 essay questions, and most of the questions have several parts.

- Many of the questions are vague, using words such as “love.”

- Many of the questions are leading – they suggest to the respondent which response she should make.
1.2 Terminology

**Target population** : The complete collection of individuals or elements we want to study.

**Sampled population** : The collection of all possible elements that might have been chosen in a sample; the population from which the sample was taken.

**Sampling unit** : The unit we actually sample. Sampling units can be the individual elements, or clusters.

**Observation unit** : The unit we take measurement from. Observation units are usually the individual elements.

**Sampling frame** : The list of sampling units.
1.3 Selection bias occurs when some part of the target population is not in the sampled population.

- A sample of convenience is often biased, since the units that are easiest to select or that are most likely to respond are usually not representative of the harder-to-select or non-responding units.

- A judgment sample – the investigator uses his or her judgment to select the specific units to be included in the sample.

- Undercoverage – failing to include all of the target population in the sampling frame.

- Overcoverage – including population units in the sampling frame that are not in the target population.

- Nonresponse – failing to obtain responses from all of the chosen sample.
1.4 Sampling steps:

1. A clear statement of objectives.
2. The population to be sampled.
3. The relevant data to be collected: define study variable(s) and population quantities.
4. Required precision of estimates.
5. The population frame: define sampling units and construct the list of the sampling units.
6. Method of selecting the sample.
7. Organization of the field work.
8. Plans for handling nonresponse.
9. Summarizing and analyzing the data: estimation procedures and other statistical techniques to be employed.
10. Writing reports.
1.5 Justifications for sampling  There are three main justifications for using sampling:

1. Sampling can provide reliable information at far less cost. With a fixed budget, performing a census is often impracticable.

2. Data can be collected more quickly, so results can be published in a timely fashion. Knowing the exact unemployment rate for the year 2005 is not very helpful if it takes two years to complete the census.

3. Estimates based on sample surveys are often more accurate than the results based on a census. This is a little surprising. A census often requires a large administrative organization and involves many persons in the data collection. Biased measurement, wrong recording, and other types of errors can be easily injected into the census. In a sample, high quality data can be obtained through well trained personnel and following up studies on nonrespondents.
2. Simple Probability Samples

2.1 Types of probability samples

- A simple random sample (SRS) of size \( n \) is taken when every possible subset of \( n \) units in the population has the same chance of being the sample.

- In a stratified random sample, the population is divided into subgroups called strata. Then SRSs are independently selected from these strata.

- In a cluster sample, observation units in the population are aggregated into larger sampling units, called clusters, which are randomly selected with all elements included.

- In a systematic sample, a starting point is chosen from a list of population members using a random number. That unit, and every \( k \)th unit thereafter, is chosen.
2.1 Types of Probability Samples

FIGURE 2.1
Examples of a simple random sample, stratified random sample, cluster sample, and systematic sample of 20 integers from the population \{1, 2, \ldots, 100\}.

Simple random sample of 20 numbers from population of 100 numbers

Stratified random sample of 20 numbers from population of 100 numbers

Cluster sample of 20 numbers from population of 100 numbers

Systematic sample of 20 numbers from population of 100 numbers

each faculty member in those departments how much time he or she spent grading homework. A systematic sample could be chosen by selecting an integer at random between 1 and 20; if the random integer is 16, say, then you would include professors in positions 16, 36, 56, and so on, in the list.

EXAMPLE 2.1
Figure 2.1 illustrates the differences among simple random, stratified, cluster, and systematic sampling for selecting a sample of 20 integers from the population \{1, 2, \ldots, 100\}. For the stratified sample, the population was divided into the 10 strata \{1, 2, \ldots, 10\}, \{11, 12, \ldots, 20\}, \ldots, \{91, 92, \ldots, 100\}, and an SRS of 2 numbers was drawn from each of the 10 strata. This ensures that each stratum is represented in the sample. For the cluster sample, the population was divided into 20 clusters \{1, 2, 3, 4, 5\}, \{6, 7, 8, 9, 10\}, \ldots, \{96, 97, 98, 99, 100\}; an SRS of 4 of these clusters was selected. For the systematic sample, the random starting point was 3, so the sample contains units 3, 8, 13, 18, and so on.

All of these methods—simple random sampling, stratified random sampling, cluster sampling, and systematic sampling—involves random selection of units to be in the sample. In an SRS, the observation units themselves are selected at random from the
population of observation units; in a stratified random sample, observation units within each stratum are randomly selected; in a cluster sample, the clusters are randomly selected from the population of all clusters. Each method is a form of probability sampling, which we discuss in the next section.

2.2 Framework for Probability Sampling

To show how probability sampling works, we need to be able to list the \( N \) units in the finite population. The finite population, or universe, of \( N \) units is denoted by the index set

\[
U = \{1, 2, \ldots, N\}. \tag{2.1}
\]

Out of this population we can choose various samples, which are subsets of \( U \). The particular sample chosen is denoted by \( S \), a subset consisting of \( n \) of the units in \( U \).

Suppose the population has four units: \( U = \{1, 2, 3, 4\} \). Six different samples of size 2 could be chosen from this population:

\[
S_1 = \{1, 2\}, \quad S_4 = \{2, 3\}, \\
S_2 = \{1, 3\}, \quad S_5 = \{2, 4\}, \\
S_3 = \{1, 4\}, \quad S_6 = \{3, 4\}.
\]

In probability sampling, each possible sample \( S \) from the population has a known probability \( P(S) \) of being chosen, and the probabilities of the possible samples sum to 1. One possible sample design for a probability sample of size 2 would have \( P(S_1) = 1/3 \), \( P(S_2) = 1/6 \), and \( P(S_6) = 1/2 \), and \( P(S_3) = P(S_4) = P(S_5) = 0 \). The probabilities \( P(S_1), P(S_2), \) and \( P(S_6) \) of the possible samples are known before the sample is drawn. One way to select the sample would be to place six labeled balls in a box; two of the balls are labeled 1, one is labeled 2, and three are labeled 6. Now choose one ball at random; if a ball labeled 6 is chosen, then \( S_6 \) is the sample.

In a probability sample, since each possible sample has a known probability of being the chosen sample, each unit in the population has a known probability of appearing in our selected sample. We calculate

\[
\pi_i = P(\text{unit } i \text{ in sample}) \tag{2.2}
\]

by summing the probabilities of all possible samples that contain unit \( i \). In probability sampling, the \( \pi_i \) are known before the survey commences, and we assume that \( \pi_i > 0 \) for every unit in the population. For the sample design described above, \( \pi_1 = P(S_1) + P(S_2) + P(S_3) = 1/2 \), \( \pi_2 = P(S_1) + P(S_4) + P(S_5) = 1/3 \), \( \pi_3 = P(S_2) + P(S_4) + P(S_6) = 2/3 \), and \( \pi_4 = P(S_3) + P(S_5) + P(S_6) = 1/2 \).

Of course, we never write all possible samples down and calculate the probability with which we would choose every possible sample—this would take far too long. But such enumeration underlies all of probability sampling. Investigators using a probability sample have much less discretion about which units are included in the sample, so using probability samples helps us avoid some of the selection biases described in Chapter 1. In a probability sample, the interviewer cannot choose to substitute a friendly looking person for the grumpy person selected to be in the sample.
by the random selection method. A forester taking a probability sample of trees cannot simply measure the trees near the road but must measure the trees designated for inclusion in the sample. Taking a probability sample is much harder than taking a convenience sample, but a probability sampling procedure guarantees that each unit in the population could appear in the sample and provides information that can be used to assess the precision of statistics calculated from the sample.

Within the framework of probability sampling, we can quantify how likely it is that our sample is a “good” one. A single probability sample is not guaranteed to be representative of the population with regard to the characteristics of interest, but we can quantify how often samples will meet some criterion of representativeness. The notion is the same as that of confidence intervals: We do not know whether the particular 95% confidence interval we construct for the mean contains the true value of the mean. We do know, however, that if the assumptions for the confidence interval procedure are valid and if we repeat the procedure over and over again, we can expect 95% of the resulting confidence intervals to contain the true value of the mean.

Let \( y_i \) be a characteristic associated with the \( i \)th unit in the population. We consider \( y_i \) to be a fixed quantity; if Farm 723 is included in the sample, then the amount of corn produced on Farm 723, \( y_{723} \), is known exactly.

**Example 2.2** To illustrate these concepts, let’s look at an artificial situation in which we know the value of \( y_i \) for each of the \( N = 8 \) units in the whole population. The index set for the population is

\[
U = \{1, 2, 3, 4, 5, 6, 7, 8\}.
\]

The values of \( y_i \) are

<table>
<thead>
<tr>
<th>( i )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_i )</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

There are 70 possible samples of size 4 that may be drawn without replacement from this population; the samples are listed in file samples.dat on the website. If the sample consisting of units \{1, 2, 3, 4\} were chosen, the corresponding values of \( y_i \) would be 1, 2, 4, and 4. The values of \( y_i \) for the sample \{2, 3, 6, 7\} are 2, 4, 7, and 7. Define \( P(S) = 1/70 \) for each distinct subset of size four from \( U \). As you will see after you read Section 2.3, this design is an SRS without replacement. Each unit is in exactly 35 of the possible samples, so \( \pi_i = 1/2 \) for \( i = 1, 2, \ldots, 8 \).

A random mechanism is used to select one of the 70 possible samples. One possible mechanism for this example, because we have listed all possible samples, is to generate a random number between 1 and 70 and select the corresponding sample. With large populations, however, the number of samples is so great that it is impractical to list all possible samples—instead, another method is used to select the sample. Methods that will give an SRS will be described in Section 2.3.

Most results in sampling rely on the sampling distribution of a statistic, the distribution of different values of the statistic obtained by the process of taking all possible samples from the population. A sampling distribution is an example of a discrete probability distribution.
Suppose we want to use a sample to estimate a population quantity, say the population total \( t = \sum_{i=1}^{N} y_i \). One estimator we might use for \( t \) is \( \hat{t}_S = N\bar{y}_S \), where \( \bar{y}_S \) is the average of the \( y_i \)'s in \( S \), the chosen sample. In our example, \( t = 40 \). If the sample \( S \) consists of units 1, 3, 5, and 6, then \( \hat{t}_S = 8 \times (1 + 4 + 7 + 7)/4 = 38 \). Since we know the whole population here, we can find \( \hat{t}_S \) for each of the 70 possible samples. The probabilities of selection for the samples give the sampling distribution of \( \hat{t} \):

\[
P(\hat{t} = k) = \sum_{S : \hat{t}_S = k} P(S).
\]

The summation is over all samples \( S \) for which \( \hat{t}_S = k \). We know the probability \( P(S) \) with which we select a sample \( S \) because we take a probability sample.

**Example 2.3** The sampling distribution of \( \hat{t} \) for the population and sampling design in Example 2.2 derives entirely from the probabilities of selection for the various samples. Four samples (\{3,4,5,6\}, \{3,4,5,7\}, \{3,4,6,7\}, and \{1,5,6,7\}) result in the estimate \( \hat{t} = 44 \), so \( P(\hat{t} = 44) = 4/70 \). For this example, we can write out the sampling distribution of \( \hat{t} \) because we know the values for the entire population.

| \( k \) | 22 | 28 | 30 | 32 | 34 | 36 | 38 | 40 | 42 | 44 | 46 | 48 | 50 | 52 | 58 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| \( P(\hat{t} = k) \) | 1/70 | 6/70 | 2/70 | 3/70 | 7/70 | 4/70 | 6/70 | 12/70 | 6/70 | 4/70 | 7/70 | 3/70 | 2/70 | 6/70 | 1/70 |

Figure 2.2 displays the sampling distribution.

The expected value of \( \hat{t} \), \( E[\hat{t}] \), is the mean of the sampling distribution of \( \hat{t} \):

\[
E[\hat{t}] = \sum_{S} \hat{t}_S P(S) = \sum_{k} k P(\hat{t} = k).
\]

**Figure 2.2**
Sampling distribution of the sample total in Example 2.3.
The expected value of the statistic is the weighted average of the possible sample values of the statistic, weighted by the probability that particular value of the statistic would occur.

The estimation bias of the estimator \( \hat{t} \) is

\[
\text{Bias} [ \hat{t} ] = E[\hat{t}] - t. \tag{2.4}
\]

If \( \text{Bias}[\hat{t}] = 0 \), we say that the estimator \( \hat{t} \) is unbiased for \( t \). For the data in Example 2.2 the expected value of \( \hat{t} \) is

\[
E[\hat{t}] = \frac{1}{70}(22) + \frac{6}{70}(28) + \cdots + \frac{1}{70}(58) = 40.
\]

Thus, the estimator is unbiased.

Note that the mathematical definition of bias in (2.4) is not the same thing as the selection or measurement bias described in Chapter 1. All indicate a systematic deviation from the population value, but from different sources. Selection bias is due to the method of selecting the sample—often, the investigator acts as though every possible sample \( S \) has the same probability of being selected, but some subsets of the population actually have a different probability of selection. With undercoverage, for example, the probability of including a unit not in the sampling frame is zero. Measurement bias means that the \( y_i \)'s are not really the quantities of interest, so although \( \hat{t} \) may be unbiased in the sense of (2.4) for \( t = \sum_{i=1}^{N} y_i \), \( t \) itself would not be the true total of interest. Estimation bias means that the estimator chosen results in bias—for example, if we used \( \hat{t}_S = \sum_{i \in S} y_i \) and did not take a census, \( \hat{t} \) would be biased. To illustrate these distinctions, suppose you wanted to estimate the average height of male actors belonging to the Screen Actors Guild. Selection bias would occur if you took a convenience sample of actors on the set—perhaps taller actors are more or less likely to be working. Measurement bias would occur if your tape measure inaccurately added 3 centimeters (cm) to each actor’s height. Estimation bias would occur if you took an SRS from the list of all actors in the Guild, but estimated mean height by the average height of the six shortest men in the sample—the sampling procedure is good, but the estimator is bad.

The variance of the sampling distribution of \( \hat{t} \) is

\[
V(\hat{t}) = E[(\hat{t} - E[\hat{t}])^2] = \sum_{\text{all possible samples } S} P(S) [\hat{t}_S - E(\hat{t})]^2. \tag{2.5}
\]

For the data in Example 2.2,

\[
V(\hat{t}) = \frac{1}{70}(22 - 40)^2 + \cdots + \frac{1}{70}(58 - 40)^2 = \frac{3840}{70} = 54.86.
\]

Because we sometimes use biased estimators, we often use the mean squared error (MSE) rather than variance to measure the accuracy of an estimator.

\[
\text{MSE}[\hat{t}] = E[(\hat{t} - t)^2]
\]

\[
= E[(\hat{t} - E[\hat{t}] + E[\hat{t}] - t)^2]
\]

\[
= E[(\hat{t} - E[\hat{t}])^2] + (E[\hat{t}] - t)^2 + 2E[(\hat{t} - E[\hat{t}])(E[\hat{t}] - t)]
\]

\[
= V(\hat{t}) + \left[ \text{Bias}(\hat{t}) \right]^2.
\]
Figure 2.3
Unbiased, precise, and accurate archers. Archer A is unbiased—the average position of all arrows is at the bull’s-eye. Archer B is precise but not unbiased—all arrows are close together but systematically away from the bull’s-eye. Archer C is accurate—all arrows are close together and near the center of the target.

Thus, an estimator \( \hat{t} \) of \( t \) is **unbiased** if \( E(\hat{t}) = t \), **precise** if \( V(\hat{t}) = E((\hat{t} - E(\hat{t}))^2) \) is small, and **accurate** if MSE(\( \hat{t} \)) = \( E((\hat{t} - t)^2) \) is small. A badly biased estimator may be precise but it will not be accurate; accuracy (MSE) is how close the estimate is to the true value, while precision (variance) measures how close estimates from different samples are to each other. Figure 2.3 illustrates these concepts.

In summary, the finite population \( \mathcal{U} \) consists of units \( \{1, 2, \ldots, N\} \) whose measured values are \( \{y_1, y_2, \ldots, y_N\} \). We select a sample \( S \) of \( n \) units from \( \mathcal{U} \) using the probabilities of selection that define the sampling design. The \( y_i \)'s are fixed but unknown quantities—unknown unless that unit happens to appear in our sample \( S \). Unless we make additional assumptions, the only information we have about the set of \( y_i \)'s in the population is in the set \( \{y_i : i \in S\} \).

You may be interested in many different population quantities from your population. Historically, however, the main impetus for developing theory for sample surveys has been estimating population means and totals. Suppose we want to estimate the total number of persons in Canada who have diabetes, or the average number of oranges produced per orange tree. The population total is

\[
t = \sum_{i=1}^{N} y_i,
\]

and the mean of the population is

\[
\bar{y}_U = \frac{1}{N} \sum_{i=1}^{N} y_i.
\]

Almost all populations exhibit some variability; for example, households have different incomes and trees have different diameters. Define the **variance** of the population values about the mean as

\[
S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_U)^2.
\]

(2.6)
2.3 Simple Random Sampling

The population standard deviation is $S = \sqrt{S^2}$.

It is sometimes helpful to have a special notation for proportions. The proportion of units having a characteristic is simply a special case of the mean, obtained by letting $y_i = 1$ if unit $i$ has the characteristic of interest, and $y_i = 0$ if unit $i$ does not have the characteristic. Let

$$p = \frac{\text{number of units with the characteristic in the population}}{N}.$$ 

**Example 2.4** For the population in Example 2.2, let

$$y_i = \begin{cases} 
1 & \text{if unit } i \text{ has the value 7} \\
0 & \text{if unit } i \text{ does not have the value 7}
\end{cases}$$

Let $\hat{p}_S = \frac{\sum_{i \in S} y_i}{4}$, the proportion of 7s in the sample. The list of all possible samples in the data file samples.dat has 5 samples with no 7s, 30 samples with exactly one 7, 30 samples with exactly two 7s, and 5 samples with three 7s. Since one of the possible samples is selected with probability 1/70, the sampling distribution of $\hat{p}$ is:

<table>
<thead>
<tr>
<th>$k$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\hat{p} = k)$</td>
<td>$\frac{5}{70}$</td>
<td>$\frac{30}{70}$</td>
<td>$\frac{30}{70}$</td>
<td>$\frac{5}{70}$</td>
</tr>
</tbody>
</table>

2.3 Simple Random Sampling

Simple random sampling is the most basic form of probability sampling, and provides the theoretical basis for the more complicated forms. There are two ways of taking a simple random sample: with replacement, in which the same unit may be included more than once in the sample, and without replacement, in which all units in the sample are distinct.

A **simple random sample with replacement** (SRSWR) of size $n$ from a population of $N$ units can be thought of as drawing $n$ independent samples of size 1. One unit is randomly selected from the population to be the first sampled unit, with probability $1/N$. Then the sampled unit is replaced in the population, and a second unit is randomly selected with probability $1/N$. This procedure is repeated until the sample has $n$ units, which may include duplicates from the population.

In finite population sampling, however, sampling the same person twice provides no additional information. We usually prefer to sample without replacement, so that the sample contains no duplicates. A **simple random sample without replacement** (SRS) of size $n$ is selected so that every possible subset of $n$ distinct units in the population has the same probability of being selected as the sample. There are $\binom{N}{n}$ possible samples (see Appendix A), and each is equally likely, so the probability of

---

1 An alternative derivation of the sampling distribution is in Exercise A.2 in Appendix A.
selecting any individual sample $S$ of $n$ units is

$$P(S) = \frac{1}{N \choose n} = \frac{n!(N-n)!}{N!}.$$ (2.7)

As a consequence of this definition, the probability that the $i$th unit appears in the sample is $\pi_i = n/N$, as shown in Section 2.8.

To take an SRS, you need a list of all observation units in the population; this list is the sampling frame. In an SRS, the sampling unit and observation unit coincide. Each unit is assigned a number, and a sample is selected so that each possible sample of size $n$ has the same chance of being the sample actually selected. This can be thought of as drawing numbers out of a hat; in practice, computer-generated pseudo-random numbers are usually used to select a sample.

One method for selecting an SRS of size $n$ from a population of size $N$ is to generate $N$ random numbers between 0 and 1, then select the units corresponding to the $n$ smallest random numbers to be the sample. For example, if $N = 10$ and $n = 4$, we generate 10 numbers between 0 and 1:

<table>
<thead>
<tr>
<th>unit $i$</th>
<th>random number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.837</td>
</tr>
<tr>
<td>2</td>
<td>0.636</td>
</tr>
<tr>
<td>3</td>
<td>0.465</td>
</tr>
<tr>
<td>4</td>
<td>0.609</td>
</tr>
<tr>
<td>5</td>
<td>0.154</td>
</tr>
<tr>
<td>6</td>
<td>0.766</td>
</tr>
<tr>
<td>7</td>
<td>0.821</td>
</tr>
<tr>
<td>8</td>
<td>0.713</td>
</tr>
<tr>
<td>9</td>
<td>0.987</td>
</tr>
<tr>
<td>10</td>
<td>0.469</td>
</tr>
</tbody>
</table>

The smallest 4 of the random numbers are 0.154, 0.465, 0.469 and 0.609, leading to the sample with units $\{3, 4, 5, 10\}$. Other methods that might be used to select an SRS are described in Example 2.5 and Exercises 21 and 29. Several survey software packages will select an SRS from a list of $N$ units; the file srsselect.sas on the website gives code for selecting an SRS using SAS PROC SURVEYSELECT.

**Example 2.5** The U.S. government conducts a Census of Agriculture every five years, collecting data on all farms (defined as any place from which $1000$ or more of agricultural products were produced and sold) in the 50 states.\(^2\) The Census of Agriculture provides data on number of farms, the total acreage devoted to farms, farm size, yield of different crops, and a wide variety of other agricultural measures for each of the $N = 3078$ counties and county-equivalents in the United States. The file agpop.dat contains the 1982, 1987, and 1992 information on the number of farms, acreage devoted to farms, number of farms with fewer than 9 acres, and number of farms with more than 1000 acres for the population.

To take an SRS of size 300 from this population, I generated 300 random numbers between 0 and 1 on the computer, multiplied each by 3078, and rounded the result up to the next highest integer. This procedure generates an SRSWR. If the population is large relative to the sample, it is likely that each unit in the sample only occurs once in the list. In this case, however, 13 of the 300 numbers were duplicates. The duplicates were discarded, and replaced with new randomly generated numbers between 1 and

\(^2\)The Census of Agriculture was formerly conducted by the U.S. Census Bureau; it is currently conducted by the U.S. National Agricultural Statistics Service (NASS). More information about the census and selected data are available on the web through the NASS material on www.fedstats.gov; also see www.agcensus.usda.gov.
2.3 Simple Random Sampling

3078 until all 300 numbers were distinct; the set of random numbers generated is in file selectrs.dat, and the data set for the SRS is in agsrs.dat.

The counties selected to be in the sample may not “feel” very random at first glance. For example, counties 2840, 2841, and 2842 are all in the sample while none of the counties between 2740 and 2787 appear. The sample contains 18% of Virginia counties, but no counties in Alaska, Arizona, Connecticut, Delaware, Hawaii, Rhode Island, Utah, or Wyoming. There is a quite natural temptation to want to “adjust” the random number list, to spread it out a bit more. If you want a random sample, you must resist this temptation. Research, beginning with Neyman (1934), has repeatedly demonstrated that purposive samples often do not represent the population on key variables. If you deliberately substitute other counties for those in the randomly generated sample, you may be able match the population on one particular characteristic such as geographic distribution; however, you will likely fail to match the population on characteristics of interests such as number of farms or average farm size. If you want to ensure that all states are represented, do not adjust your randomly selected sample purposively but take a stratified sample (to be discussed in Chapter 3).

Let’s look at the variable acres92, the number of acres devoted to farms in 1992. A small number of counties in the population are missing that value—in some cases, the data are withheld to prevent disclosing data on individual farms. Thus we first check to see the extent of the missing data in our sample. Fortunately, our sample has no missing data (Exercise 23 tells how likely such an occurrence is). Figure 2.4 displays a histogram of the acreage devoted to farms in each of the 300 counties.

For estimating the population mean \( \bar{y}_U \) from an SRS, we use the sample mean

\[
\bar{y}_S = \frac{1}{n} \sum_{i \in S} y_i.
\] (2.8)
In the following, we use $\bar{y}$ to refer to the sample mean and drop the subscript $S$ unless it is needed for clarity. As will be shown in Section 2.8, $\bar{y}$ is an unbiased estimator of the population mean $y_U$, and the variance of $\bar{y}$ is

$$V(\bar{y}) = \frac{S^2}{n} \left( 1 - \frac{n}{N} \right)$$

(2.9)

for $S^2$ defined in (2.6). The variance $V(\bar{y})$ measures the variability among estimates of $y_U$ from different samples.

The factor $(1 - n/N)$ is called the finite population correction (fpc). Intuitively, we make this correction because with small populations the greater our sampling fraction $n/N$, the more information we have about the population and thus the smaller the variance. If $N = 10$ and we sample all 10 observations, we would expect the variance of $\bar{y}$ to be 0 (which it is). If $N = 10$, there is only one possible sample $S$ of size 10 without replacement, with $\bar{y}_S = \bar{y}_U$, so there is no variability due to taking a sample. For a census, the fpc, and hence $V(\bar{y})$, is 0. When the sampling fraction $n/N$ is large in an SRS without replacement, the sample is closer to a census, which has no sampling variability.

For most samples that are taken from extremely large populations, the fpc is approximately 1. For large populations it is the size of the sample taken, not the percentage of the population sampled, that determines the precision of the estimator: If your soup is well stirred, you need to taste only one or two spoonfuls to check the seasoning, whether you have made 1 liter or 20 liters of soup. A sample of size 100 from a population of 100,000 units has almost the same precision as a sample of size 100 from a population of 100 million units:

$$V[\bar{y}] = \frac{S^2}{100} \cdot \frac{99,900}{100,000} = \frac{S^2}{100} \cdot (0.999)$$

for $N = 100,000$

$$V[\bar{y}] = \frac{S^2}{100} \cdot \frac{99,999,900}{100,000,000} = \frac{S^2}{100} \cdot (0.999999)$$

for $N = 100,000,000$

The population variance $S^2$, which depends on the values for the entire population, is in general unknown. We estimate it by the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y})^2.$$  

(2.10)

An unbiased estimator of the variance of $\bar{y}$ is (see Section 2.8)

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{s^2}{n}.$$  

(2.11)

The standard error (SE) is the square root of the estimated variance of $\bar{y}$:

$$\text{SE}(\bar{y}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2}{n}}.$$  

(2.12)

The population standard deviation is often related to the mean. A population of trees might have a mean height of 10 meters (m) and standard deviation of one m. A population of small cacti, however, with a mean height of 10 cm, might have a standard deviation of 1 cm. The coefficient of variation (CV) of the estimator $\bar{y}$ is a

...
measure of relative variability, which may be defined when $\bar{y}_U \neq 0$ as:

$$\text{CV}(\bar{y}) = \frac{\sqrt{V(\bar{y})}}{E(\bar{y})} = \sqrt{1 - \frac{n}{N} \frac{S}{\sqrt{n\bar{y}_U}}}. \tag{2.13}$$

If tree height is measured in meters, then $\bar{y}_U$ and $S$ are also in meters. The CV does not depend on the unit of measurement: In this example, the trees and the cacti have the same CV. We can estimate the CV of an estimator using the standard error divided by the mean (defined only when the mean is nonzero): In an SRS,

$$\hat{\text{CV}}(\bar{y}) = \frac{\text{SE}[\bar{y}]}{\bar{y}} = \sqrt{1 - \frac{n}{N} \frac{s}{\sqrt{n\bar{y}}}}. \tag{2.14}$$

The estimated CV is thus the standard error expressed as a percentage of the mean.

All these results apply to the estimation of a population total, $t$, since

$$t = \sum_{i=1}^{N} y_i = N\bar{y}_U.$$ 

To estimate $t$, we use the unbiased estimator

$$\hat{t} = N\bar{y}. \tag{2.15}$$

Then, from (2.9),

$$V(\hat{t}) = N^2V(\bar{y}) = N^2\left(1 - \frac{n}{N}\right)s^2 \tag{2.16}$$

and

$$\hat{V}(\hat{t}) = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}. \tag{2.17}$$

Note that $\text{CV}(\hat{t}) = \sqrt{V(\hat{t})}/E(\hat{t})$ is the same as $\text{CV}(\bar{y})$.

**Example 2.6** For the data in Example 2.5, $N = 3078$ and $n = 300$, so the sampling fraction is $300/3078 = 0.097$. The sample statistics are $\bar{y} = 297,897$, $s = 344,551.9$, and $\hat{t} = N\bar{y} = 916,927,110$. Standard errors are

$$\text{SE}[\bar{y}] = \sqrt{\frac{s^2}{n} \left(1 - \frac{300}{3078}\right)} = 18,898.434428$$

and

$$\text{SE}[\hat{t}] = (3078)(18,898.434428) = 58,169,381,$$
and the estimated coefficient of variation is
\[ \hat{CV}[\hat{y}] = \frac{SE[\hat{y}]}{\hat{y}} \]
\[ = \frac{18,898.434428}{297,897} \]
\[ = 0.06344. \]

Since these data are so highly skewed, we should also report the median number of farm acres in a county, which is 196,717.

We might also want to estimate the proportion of counties in Example 2.5 with fewer than 200,000 acres in farms. Since estimating a proportion is a special case of estimating a mean, the results in (2.8)–(2.14) hold for proportions as well, and they take a simple form. Suppose we want to estimate the proportion of units in the population that have some characteristic—call this proportion \( p \). Define \( y_i \) to be 1 if the unit has the characteristic and to be 0 if the unit does not have that characteristic. Then \( p = \frac{\sum_{i=1}^{N} y_i}{N} = \bar{y}_U \), and \( p \) is estimated by \( \hat{p} = \bar{y} \). Consequently, \( \hat{p} \) is an unbiased estimator of \( p \). For the response \( y_i \), taking on values 0 or 1,

\[ S^2 = \frac{\sum_{i=1}^{N} (y_i - p)^2}{N - 1} = \frac{\sum_{i=1}^{N} y_i^2 - 2p \sum_{i=1}^{N} y_i + Np^2}{N - 1} = \frac{N}{N - 1} p(1 - p). \]

Thus, (2.9) implies that
\[ V(\hat{p}) = \left( \frac{N - n}{N - 1} \right) \frac{p(1 - p)}{n}. \]

(2.18)

Also,

\[ s^2 = \frac{1}{n - 1} \sum_{i \in S} (y_i - \hat{p})^2 = \frac{n}{n - 1} \hat{p}(1 - \hat{p}), \]

so from (2.11),
\[ \hat{V}(\hat{p}) = \left( 1 - \frac{n}{N} \right) \frac{\hat{p}(1 - \hat{p})}{n - 1}. \]

(2.19)

**Example 2.7** For the sample described in Example 2.5, the estimated proportion of counties with fewer than 200,000 acres in farms is
\[ \hat{p} = \frac{153}{300} = 0.51 \]

with standard error
\[ SE(\hat{p}) = \sqrt{ \left( 1 - \frac{300}{3078} \right) \frac{(0.51)(0.49)}{299} } = 0.0275. \]
Note: In an SRS, \( \pi_i = n/N \) for all units \( i = 1, \ldots, N \). However, many other probability sampling designs also have \( \pi_i = n/N \) for all units but are not SRSs. To have an SRS, it is not sufficient for every individual to have the same probability of being in the sample; in addition, every possible sample of size \( n \) must have the same probability \( \frac{1}{\binom{N}{n}} \) of being the sample selected, as defined in (2.7). The cluster sampling design in Example 2.1, in which the population of 100 integers is divided into 20 clusters \{1, 2, 3, 4, 5\}, \{6, 7, 8, 9, 10\}, \ldots \{96, 97, 98, 99, 100\} and an SRS of 4 of these clusters selected, has \( \pi_i = 20/100 \) for each unit in the population but is not an SRS of size 20 because different possible samples of size 20 have different probabilities of being selected. To see this, let’s look at two particular subsets of \{1, 2, \ldots, 100\}. Let \( S_1 \) be the cluster sample depicted in the third panel of Figure 2.1, with
\[
S_1 = \{1, 2, 3, 4, 5, 46, 47, 48, 49, 50, 61, 62, 63, 64, 65, 81, 82, 83, 84, 85\},
\]
and let
\[
S_2 = \{1, 6, 11, 16, 21, 26, 31, 36, 41, 46, 51, 56, 61, 66, 71, 76, 81, 86, 91, 96\}.
\]
The cluster sampling design specifies taking an SRS of 4 of the 20 clusters, so 
\[
P(S_1) = \frac{1}{\binom{20}{4}} = 4!(20 - 4)!/20! = 1/4845.
\]
Sample \( S_2 \) cannot occur under this design, however, so \( P(S_2) = 0 \). An SRS with \( n = 20 \) from a population with \( N = 100 \) would have
\[
P(S) = \frac{1}{\binom{100}{20}} = \frac{20!(100 - 20)!}{100!} = \frac{1}{5.359834 \times 10^{20}}
\]
for every subset \( S \) of size 20 from the population \{1, 2, \ldots, 100\}.

## 2.4 Sampling Weights

In (2.2), we defined \( \pi_i \) to be the probability that unit \( i \) is included in the sample. In probability sampling, these inclusion probabilities are used to calculate point estimates such as \( \hat{t} \) and \( \bar{y} \). Define the sampling weight, for any sampling design, to be the reciprocal of the inclusion probability:
\[
w_i = \frac{1}{\pi_i}.
\]
(2.20)
The sampling weight of unit \( i \) can be interpreted as the number of population units represented by unit \( i \).

In an SRS, each unit has inclusion probability \( \pi_i = n/N \); consequently, all sampling weights are the same with \( w_i = 1/\pi_i = N/n \). We can thus think of every unit in the sample as representing itself plus \( N/n - 1 \) of the unsampled units in the population. Note that for an SRS,
\[
\sum_{i \in S} w_i = \sum_{i \in S} \frac{N}{n} = N,
\]
Chapter 2: Simple Probability Samples

\[ \sum_{i \in S} w_i y_i = \frac{N}{n} \sum_{i \in S} y_i = \hat{t}, \]

and

\[ \sum_{i \in S} \frac{w_i y_i}{\sum_{i \in S} w_i} = \frac{\hat{t}}{N} = \bar{y}. \]

All weights are the same in an SRS—that is, every unit in the sample represents the same number of units, \( N/n \), in the population. We call such a sample, in which every unit has the same sampling weight, a self-weighting sample.

**Example 2.8** Let’s look at the sampling weights for the sample described in Example 2.5. Here, \( N = 3078 \) and \( n = 300 \), so the sampling weight is \( w_i = 3078/300 = 10.26 \) for each unit in the sample. The first county in the data file agsrs.dat, Coffee County, Alabama, thus represents itself and 9.26 counties from the 2778 counties not included in the sample. We can create a column of sampling weights as follows:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>County</td>
<td>State</td>
<td>acres92</td>
<td>weight</td>
<td>weight*acres92</td>
</tr>
<tr>
<td>Coffee County</td>
<td>AL</td>
<td>175,209</td>
<td>10.26</td>
<td>1,797,644.34</td>
</tr>
<tr>
<td>Colbert County</td>
<td>AL</td>
<td>138,135</td>
<td>10.26</td>
<td>1,417,265.10</td>
</tr>
<tr>
<td>Lamar County</td>
<td>AL</td>
<td>56,102</td>
<td>10.26</td>
<td>575,606.52</td>
</tr>
<tr>
<td>Marengo County</td>
<td>AL</td>
<td>199,117</td>
<td>10.26</td>
<td>2,042,940.42</td>
</tr>
<tr>
<td>Marion County</td>
<td>AL</td>
<td>89,228</td>
<td>10.26</td>
<td>915,479.28</td>
</tr>
<tr>
<td>Tuscaloosa County</td>
<td>AL</td>
<td>96,194</td>
<td>10.26</td>
<td>986,950.44</td>
</tr>
<tr>
<td>Columbia County</td>
<td>AR</td>
<td>57,253</td>
<td>10.26</td>
<td>587,415.78</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Pleasants County</td>
<td>WV</td>
<td>15,650</td>
<td>10.26</td>
<td>160,569.00</td>
</tr>
<tr>
<td>Putnam County</td>
<td>WV</td>
<td>55,827</td>
<td>10.26</td>
<td>572,785.02</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td>89,369,114</td>
<td>3078</td>
<td>916,927,109.60</td>
</tr>
</tbody>
</table>

The last column is formed by multiplying columns C and D, so the entries are \( w_i y_i \). We see that \( \sum_{i \in S} w_i y_i = 916,927,110 \), which is the same value we obtained for the estimated population total in Example 2.5.

### 2.5 Confidence Intervals

When you take a sample survey, it is not sufficient to simply report the average height of trees or the sample proportion of voters who intend to vote for Candidate B in the next election. You also need to give an indication of how accurate your estimates are. In statistics, **confidence intervals** (CIs) are used to indicate the accuracy of an estimate.
TABLE 2.1  
Confidence intervals for possible samples from small population

<table>
<thead>
<tr>
<th>Sample $\mathcal{S}$</th>
<th>$y_i, i \in \mathcal{S}$</th>
<th>$\hat{t}_S$</th>
<th>$s_S$</th>
<th>CI$(\mathcal{S})$</th>
<th>$u(\mathcal{S})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1, 2, 3, 4}</td>
<td>1, 2, 4, 4</td>
<td>22</td>
<td>1.50</td>
<td>[16.00, 28.00]</td>
<td>0</td>
</tr>
<tr>
<td>{1, 2, 3, 5}</td>
<td>1, 2, 4, 7</td>
<td>28</td>
<td>2.65</td>
<td>[17.42, 38.58]</td>
<td>0</td>
</tr>
<tr>
<td>{1, 2, 3, 6}</td>
<td>1, 2, 4, 7</td>
<td>28</td>
<td>2.65</td>
<td>[17.42, 38.58]</td>
<td>0</td>
</tr>
<tr>
<td>{1, 2, 3, 7}</td>
<td>1, 2, 4, 7</td>
<td>28</td>
<td>2.65</td>
<td>[17.42, 38.58]</td>
<td>0</td>
</tr>
<tr>
<td>{1, 2, 3, 8}</td>
<td>1, 2, 4, 8</td>
<td>30</td>
<td>3.10</td>
<td>[17.62, 42.38]</td>
<td>1</td>
</tr>
<tr>
<td>{1, 2, 4, 5}</td>
<td>1, 2, 4, 7</td>
<td>28</td>
<td>2.65</td>
<td>[17.42, 38.58]</td>
<td>0</td>
</tr>
<tr>
<td>{1, 2, 4, 6}</td>
<td>1, 2, 4, 7</td>
<td>28</td>
<td>2.65</td>
<td>[17.42, 38.58]</td>
<td>0</td>
</tr>
<tr>
<td>{1, 2, 4, 7}</td>
<td>1, 2, 4, 7</td>
<td>28</td>
<td>2.65</td>
<td>[17.42, 38.58]</td>
<td>0</td>
</tr>
<tr>
<td>{1, 2, 4, 8}</td>
<td>1, 2, 4, 8</td>
<td>30</td>
<td>3.10</td>
<td>[17.62, 42.38]</td>
<td>1</td>
</tr>
<tr>
<td>{1, 2, 5, 6}</td>
<td>1, 2, 4, 7</td>
<td>28</td>
<td>2.65</td>
<td>[17.42, 38.58]</td>
<td>0</td>
</tr>
<tr>
<td>{1, 2, 5, 6}</td>
<td>1, 2, 4, 7</td>
<td>28</td>
<td>2.65</td>
<td>[17.42, 38.58]</td>
<td>0</td>
</tr>
<tr>
<td>{1, 2, 5, 6}</td>
<td>1, 2, 4, 7</td>
<td>28</td>
<td>2.65</td>
<td>[17.42, 38.58]</td>
<td>0</td>
</tr>
<tr>
<td>{1, 2, 5, 6}</td>
<td>1, 2, 4, 7</td>
<td>28</td>
<td>2.65</td>
<td>[17.42, 38.58]</td>
<td>0</td>
</tr>
<tr>
<td>{2, 3, 4, 8}</td>
<td>2, 4, 4, 8</td>
<td>36</td>
<td>2.52</td>
<td>[25.93, 46.07]</td>
<td>1</td>
</tr>
<tr>
<td>{2, 3, 5, 6}</td>
<td>2, 4, 7, 7</td>
<td>40</td>
<td>2.45</td>
<td>[30.20, 49.80]</td>
<td>1</td>
</tr>
<tr>
<td>{2, 3, 5, 7}</td>
<td>2, 4, 7, 7</td>
<td>40</td>
<td>2.45</td>
<td>[30.20, 49.80]</td>
<td>1</td>
</tr>
<tr>
<td>{2, 3, 5, 8}</td>
<td>2, 4, 7, 8</td>
<td>42</td>
<td>2.75</td>
<td>[30.98, 53.02]</td>
<td>1</td>
</tr>
<tr>
<td>{2, 3, 6, 7}</td>
<td>2, 4, 7, 7</td>
<td>40</td>
<td>2.45</td>
<td>[30.20, 49.80]</td>
<td>1</td>
</tr>
<tr>
<td>{4, 5, 6, 8}</td>
<td>4, 7, 7, 8</td>
<td>52</td>
<td>1.73</td>
<td>[45.07, 58.93]</td>
<td>0</td>
</tr>
<tr>
<td>{4, 5, 7, 8}</td>
<td>4, 7, 7, 8</td>
<td>52</td>
<td>1.73</td>
<td>[45.07, 58.93]</td>
<td>0</td>
</tr>
<tr>
<td>{4, 6, 7, 8}</td>
<td>4, 7, 7, 8</td>
<td>52</td>
<td>1.73</td>
<td>[45.07, 58.93]</td>
<td>0</td>
</tr>
<tr>
<td>{5, 6, 7, 8}</td>
<td>7, 7, 7, 8</td>
<td>58</td>
<td>0.50</td>
<td>[56.00, 60.00]</td>
<td>0</td>
</tr>
</tbody>
</table>

superpopulation, and so on until the superpopulations are as large as we could wish. Our population is embedded in a series of increasing finite populations. This embedding can give us properties such as consistency and asymptotic normality. One can imagine the superpopulations as “alternative universes” in a science fiction sense—what might have happened if circumstances were slightly different.

Hájek (1960) proves a central limit theorem for simple random sampling without replacement (also see Lehmann, 1999, Sections 2.8 and 4.4, for a derivation). In practical terms, Hájek’s theorem says that if certain technical conditions hold and if $n, N,$ and $N - n$ are all “sufficiently large,” then the sampling distribution of

$$
\bar{y} - \tilde{y}_U \sqrt{\frac{1}{N - n}} \mathcal{S} \sqrt{n}
$$

is approximately normal (Gaussian) with mean 0 and variance 1. A large-sample $100(1 - \alpha)\%$ CI for the population mean is

$$
\left[ \bar{y} - z_{\alpha/2} \sqrt{\frac{1 - n}{N} \frac{S}{\sqrt{n}}}, \bar{y} + z_{\alpha/2} \sqrt{\frac{1 - n}{N} \frac{S}{\sqrt{n}}} \right],
$$

(2.21)
where \( z_{\alpha/2} \) is the \((1 - \alpha/2)\)th percentile of the standard normal distribution. In simple random sampling without replacement, 95\% of the possible samples that could be chosen will give a 95\% CI for \( \bar{y}_U \) that contains the true value of \( \bar{y}_U \). Usually, \( S \) is unknown, so in large samples \( s \) is substituted for \( S \) with little change in the approximation; the large-sample CI is

\[
[\bar{y} - z_{\alpha/2}SE(\bar{y}), \bar{y} + z_{\alpha/2}SE(\bar{y})].
\]

In practice, we often substitute \( t_{\alpha/2, n-1} \), the \((1 - \alpha/2)\)th percentile of a \( t \) distribution with \( n - 1 \) degrees of freedom, for \( z_{\alpha/2} \). For large samples, \( t_{\alpha/2, n-1} \approx z_{\alpha/2} \). In smaller samples, using \( t_{\alpha/2, n-1} \) instead of \( z_{\alpha/2} \) produces a wider CI. Most software packages use the following CI for the population mean from an SRS:

\[
[\bar{y} - t_{\alpha/2, n-1} \sqrt{1 - \frac{n}{N}} \frac{s}{\sqrt{n}} \bar{y}, \bar{y} + t_{\alpha/2, n-1} \sqrt{1 - \frac{n}{N}} \frac{s}{\sqrt{n}}], \tag{2.22}
\]

The imprecise term “sufficiently large” occurs in the central limit theorem because the adequacy of the normal approximation depends on \( n \) and on how closely the population \( \{y_i, i = 1, \ldots, N\} \) resembles a population generated from the normal distribution. The “magic number” of \( n = 30 \), often cited in introductory statistics books as a sample size that is “sufficiently large” for the central limit theorem to apply, often does not suffice in finite population sampling problems. Many populations we sample are highly skewed—we may measure income, number of acres on a farm that are devoted to corn, or the concentration of mercury in Minnesota lakes. For all of these examples, we expect most of the observations to be relatively small, but a few to be very, very large, so that a smoothed histogram of the entire population would look like this:

![Histogram](image)

Thinking of observations as generated from some distribution is useful in deciding whether it is safe to use the central limit theorem. If you can think of the generating distribution as being somewhat close to normal, it is probably safe to use the central limit theorem with a sample size as small as 50. If the sample size is too small and the sampling distribution of \( \bar{y} \) not approximately normal, we need to use another method, relying on distributional assumptions, to obtain a CI for \( \bar{y}_U \). Such methods rely on a model-based perspective for sampling (Section 2.9).
The mean for the $i$th cluster is $\bar{Y}_i = M^{-1} \sum_{j=1}^{M} y_{ij}$, and the variance for the $i$th cluster is $S_i^2 = (M - 1)^{-1} \sum_{j=1}^{M} (y_{ij} - \bar{Y}_i)^2$.

**One-stage cluster sampling:** Take $n$ clusters (denoted by $s$) using simple random sampling without replacement, and all elements in the selected clusters are observed. The sample mean (per element) is given by

$$\bar{y} = \frac{1}{nM} \sum_{i \in s} \sum_{j=1}^{M} y_{ij} = \frac{1}{n} \sum_{i \in s} \bar{Y}_i.$$ 

**Result 2.6** Under one-stage cluster sampling with clusters sampled using SRSWOR,

(i) $E(\bar{y}) = \bar{Y}$.

(ii) $V(\bar{y}) = (1 - \frac{n}{N}) \frac{S_M^2}{n}$, where $S_M^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y})^2$.

(iii) $v(\bar{y}) = (1 - \frac{n}{N}) \frac{1}{n} \frac{1}{n-1} \sum_{i \in s} (\bar{Y}_i - \bar{y})^2$ is an unbiased estimator for $V(\bar{y})$.

When cluster sizes are not all equal, complications will arise. When $M_i$’s are all known, simple solutions exist, otherwise a ratio type estimator will have to be used. It is also interesting to note that systematic sampling is a special case of one-stage cluster sampling.

### 2.6 Sample size determination

In planning a survey, one needs to know how big a sample he should draw. The answer to this question depends on how accurate he wants the estimate to be. We assume the sampling scheme is SRSWOR.

1. **Precision specified by absolute tolerable error**

   The surveyor can specify the margin of error, $e$, such that

   $$P(|\bar{y} - \bar{Y}| > e) \leq \alpha$$

   for a chosen value of $\alpha$, usually taken as 0.05. Approximately we have

   $$e = z_{\alpha/2} \sqrt{\frac{1 - \frac{n}{N}}{n} \frac{S}{\sqrt{n}}}.$$
2.6. SAMPLE SIZE DETERMINATION

Solving for \( n \), we have

\[
n = \frac{z_{\alpha/2}^2 S^2}{e^2 + z_{\alpha/2}^2 S^2/N} = \frac{n_0}{1 + n_0/N}
\]

where \( n_0 = z_{\alpha/2}^2 S^2 / e^2 \).

2. Precision specified by relative tolerable error

The precision is often specified by a relative tolerable error, \( e \).

\[
P \left( \frac{|\bar{y} - \bar{Y}|}{|\bar{Y}|} > e \right) \leq \alpha
\]

The required \( n \) is given by

\[
n = \frac{z_{\alpha/2}^2 S^2}{e^2 \bar{Y}^2 + z_{\alpha/2}^2 S^2/N} = \frac{n_0^*}{1 + n_0^*/N}.
\]

Where \( n_0^* = z_{\alpha/2}^2 (CV)^2 / e^2 \), and \( CV = S/\bar{Y} \) is the coefficient of variation.

3. Sample size for estimating proportions

The absolute tolerable error is often used, \( P(|p - P| > e) \leq \alpha \), and the common choice of \( e \) and \( \alpha \) are 3% and 0.05. Also note that \( S^2 \doteq P(1 - P) \), \( 0 \leq P \leq 1 \) implies \( S^2 \leq 1/4 \). The largest value of required sample size \( n \) occurs at \( P = 1/2 \).

Sample size determination requires the knowledge of \( S^2 \) or \( CV \). There are two ways to obtain information on these.

(a) Historical data. Quite often there were similar studies conducted previously, and information from these studies can be used to get approximate values for \( S^2 \) or \( CV \).

(b) A pilot survey. Use a small portion of the available resource to conduct a small scale pilot survey before the formal one to obtain information about \( S^2 \) or \( CV \).

Other methods are often ad hoc. For example, if a population has a range of 100. That is, the largest value minus the smallest value is no more than 100. Then a conventional estimate of \( S \) is 100/4. This example is applicable when the age is the study variable.
may be expressed as

\[ P \left( \left| \frac{\bar{y} - \bar{y}_U}{\bar{y}_U} \right| \leq r \right) = 1 - \alpha. \]

**Find an Equation** The simplest equation relating the precision and sample size comes from the confidence intervals in the previous section. To obtain absolute precision \( e \), find a value of \( n \) that satisfies

\[ e = z_{\alpha/2} \sqrt{\left( 1 - \frac{n}{N} \right) \frac{S}{\sqrt{n}}}. \]

To solve this equation for \( n \), we first find the sample size \( n_0 \) that we would use for an SRSWR:

\[ n_0 = \left( \frac{z_{\alpha/2} S}{e} \right)^2. \quad (2.24) \]

Then (see Exercise 9) the desired sample size is

\[ n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{\frac{z_{\alpha/2}^2 S^2}{e^2 + \frac{z_{\alpha/2}^2 S^2}{N}}}{1 + \frac{n_0}{N}}. \quad (2.25) \]

Of course, if \( n_0 \geq N \) we simply take a census with \( n = N \).

In surveys in which one of the main responses of interest is a proportion, it is often easiest to use that response in setting the sample size. For large populations, \( S^2 \approx p(1 - p) \), which attains its maximal value when \( p = 1/2 \). So using \( n_0 = 1.96^2/(4e^2) \) will result in a 95% CI with width at most 2\( e \).

To calculate a sample size to obtain a specified relative precision, substitute \( r\bar{y}_U \) for \( e \) in (2.24) and (2.25). This results in sample size

\[ n = \frac{\frac{z_{\alpha/2}^2 S^2}{(r\bar{y}_U)^2 + \frac{z_{\alpha/2}^2 S^2}{N}}}{1 + \frac{n_0}{N}}. \quad (2.26) \]

To achieve a specified relative precision, the sample size may be determined using only the ratio \( S/\bar{y}_U \), the CV for a sample of size 1.

**Example 2.11** Suppose we want to estimate the proportion of recipes in the *Better Homes & Gardens New Cook Book* that do not involve animal products. We plan to take an SRS of the \( N = 1251 \) test kitchen-tested recipes, and want to use a 95% CI with margin of error 0.03. Then,

\[ n_0 = \frac{(1.96)^2 \left( \frac{1}{2} \right) \left( 1 - \frac{1}{2} \right)}{(0.03)^2} \approx 1067. \]
The sample size ignoring the fpc is large compared with the population size, so in this case we would make the fpc adjustment and use
\[ n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{1067}{1 + \frac{1067}{1251}} = 576. \tag{2.11} \]

In Example 2.11, the fpc makes a difference in the sample size because \( N \) is only 1251. If \( N \) is large, however, typically \( n_0/N \) will be very small so that for large populations we usually have \( n \approx n_0 \). Thus, we need approximately the same sample size for any large population—whether that population has 10 million or 1 billion or 100 billion units.

**Example 2.12** Many public opinion polls specify using a sample size of about 1100. That number comes from rounding the value of \( n_0 \) in Example 2.11 up to the next hundred, and then noting that the population size is so large relative to the sample that the fpc should be ignored. For large populations, it is the size of the sample, not the proportion of the population that is sampled, that determines the precision. ■

**Estimate unknown quantities.** When interested in a proportion, we can use 1/4 as an upper bound for \( S^2 \). For other quantities, \( S^2 \) must be estimated or guessed at. Some methods for estimating \( S^2 \) include:

1. Use sample quantities obtained when pretesting your survey. This is probably the best method, as your pretest should be similar to the survey you take. A **pilot sample**, a small sample taken to provide information and guidance for the design of the main survey, can be used to estimate quantities needed for setting the sample size.

2. Use previous studies or data available in the literature. You are rarely the first person in the world to study anything related to your investigation. You may be able to find estimates of variances that have been published in related studies, and use these as a starting point for estimating your sample size. But you have no control over the quality or design of those studies, and their estimates may be unreliable or may not apply for your study. In addition, estimates may change over time and vary in different geographic locations.

   Sometimes you can use the CV for a sample of size 1, the ratio of the standard deviation to the mean, in obtaining estimates of variability. The CV of a quantity is a measure of relative error, and tends to be more stable over time and location than the variance. If we take a random sample of houses for sale in the United States today, we will find that the variability in price will be much greater than if we had taken a similar survey in 1930. But the average price of a house has also increased from 1930 to today. We would probably find that the CV today is close to the CV in 1930.

3. If nothing else is available, guess the variance. Sometimes a hypothesized distribution of the data will give us information about the variance. For example, if you believe the population to be normally distributed, you may not know what the variance is, but you may have an idea of the range of the data. You could then estimate
2.6 Sample Size Estimation

2.6 Sample Size Estimation

FIGURE 2.5
Plot of $t_{0.025,n−1}s/\sqrt{n}$ vs. $n$, for two possible values of the standard deviation $s$.

$S$ by range/4 or range/6, as approximately 95% of values from a normal population are within 2 standard deviations of the mean, and 99.7% of the values are within 3 standard deviations of the mean.

EXAMPLE 2.13 Before taking the sample of size 300 in Example 2.5, we took a pilot sample of size 30 from the population. One county in the pilot sample of size 30 was missing the value of $\text{acres92}$; the sample standard deviation of the remaining 29 observations was 519,085. Using this value, and a desired margin of error of 60,000,

$$n_0 = (1.96)^2 \frac{519,085^2}{60,000^2} = 288.$$

We took a sample of size 300 in case the estimated standard deviation from the pilot sample is too low. Also, we ignored the fpc in the sample size calculations; in most populations, the fpc will have little effect on the sample size.

You may also view possible consequences of different sample sizes graphically. Figure 2.5 shows the value of $t_{0.025,n−1}s/\sqrt{n}$, for a range of sample sizes between 50 and 700, and for two possible values of the standard deviation $s$. The plot shows that if we ignore the fpc and if the standard deviation is about 500,000, a sample of size 300 will give a margin of error of about 60,000.

Determining the sample size is one of the early steps that must be taken in an investigation, and no magic formula will tell you the perfect sample size for your
to him at the White House. This systematic sample most likely behaved much like a random sample. Note that Kennedy was well aware that the letters he read, while representative of letters written to the White House, were not at all representative of public opinion.

Systematic sampling does not necessarily give a representative sample, though, if the listing of population units is in some periodic or cyclical order. If male and female names alternate in the list, for example, and \( k \) is even, the systematic sample will contain either all men or all women—this cannot be considered a representative sample. In ecological surveys done on agricultural land, a ridge-and-furrow topography may be present that would lead to a periodic pattern of vegetation. If a systematic sampling scheme follows the same cycle, the sample will not behave like an SRS.

On the other hand, some populations are in increasing or decreasing order. A list of accounts receivable may be ordered from largest amount to smallest amount. In this case, estimates from the systematic sample may have smaller (but unestimable) variance than comparable estimates from the SRS. A systematic sample from an ordered list of accounts receivable is forced to contain some large amounts and some small amounts. It is possible for an SRS to contain all small amounts or all large amounts, so there may be more variability among the sample means of all possible SRSs than there is among the sample means of all possible systematic samples.

In systematic sampling, we must still have a sampling frame and be careful when defining the target population. Sampling every 20th student to enter the library will not give a representative sample of the student body. Sampling every 10th person exiting an airplane, though, will probably give a representative sample of the persons on that flight. The sampling frame for the airplane passengers is not written down, but it exists all the same.

## 2.8 Randomization Theory Results for Simple Random Sampling

In this section, we show that \( \bar{y} \) is an unbiased estimator of \( \bar{y}_U \). We also calculate the variance of \( \bar{y} \) given in (2.9), and show that the estimator in (2.11) is unbiased over repeated sampling.

No distributional assumptions are made about the \( y_i \)'s in order to ascertain that \( \bar{y} \) is unbiased for estimating \( \bar{y}_U \). We do not, for instance, assume that the \( y_i \)'s are normally distributed with mean \( \mu \). In the randomization theory (also called design-based) approach to sampling, the \( y_i \)'s are considered to be fixed but unknown numbers—the random variables used in randomization theory inference indicate which population units are in the sample.

Let’s see how the randomization theory works for deriving properties of the sample mean in simple random sampling. As in Cornfield (1944), define

\[
Z_i = \begin{cases} 
1 & \text{if unit } i \text{ is in the sample} \\
0 & \text{otherwise}
\end{cases} \quad (2.27)
\]

\(^3\)An asterisk (*) indicates a section, chapter, or exercise that requires more mathematical background.
Then
\[
\bar{y} = \sum_{i \in S} \frac{y_i}{n} = \sum_{i=1}^{N} Z_i \frac{y_i}{n}. \tag{2.28}
\]

The $Z_i$’s are the only random variables in (2.28) because, according to randomization theory, the $y_i$’s are fixed quantities. When we choose an SRS of $n$ units out of the $N$ units in the population, \{\(Z_1, \ldots, Z_N\)\} are identically distributed Bernoulli random variables with
\[
\pi_i = P(Z_i = 1) = P(\text{select unit } i \text{ in sample}) = \frac{n}{N} \tag{2.29}
\]
and
\[
P(Z_i = 0) = 1 - \pi_i = 1 - \frac{n}{N}.
\]

The probability in (2.29) follows from the definition of an SRS. To see this, note that if unit $i$ is in the sample, then the other $n-1$ units in the sample must be chosen from the other $N-1$ units in the population. A total of \(\binom{N-1}{n-1}\) possible samples of size $n-1$ may be drawn from a population of size $N-1$, so
\[
P(Z_i = 1) = \frac{\text{number of samples including unit } i}{\text{number of possible samples}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.
\]

As a consequence of (2.29),
\[
E[Z_i] = E[Z_i^2] = \frac{n}{N}
\]
and
\[
E[\bar{y}] = E \left[ \sum_{i=1}^{N} Z_i \frac{y_i}{n} \right] = \sum_{i=1}^{N} E[Z_i] \frac{y_i}{n} = \sum_{i=1}^{N} \frac{n}{N} \frac{y_i}{n} = \sum_{i=1}^{N} \frac{y_i}{N} = \bar{y}_U. \tag{2.30}
\]

This shows that $\bar{y}$ is an unbiased estimator of $\bar{y}_U$. Note that in (2.30), the random variables are $Z_1, \ldots, Z_N$; $y_1, \ldots, y_N$ are treated as constants.

The variance of $\bar{y}$ is also calculated using properties of the random variables $Z_1, \ldots, Z_N$. Note that
\[
V(Z_i) = E[Z_i^2] - (E[Z_i])^2 = \frac{n}{N} - \left( \frac{n}{N} \right)^2 = \frac{n}{N} \left( 1 - \frac{n}{N} \right).
\]

For $i \neq j$,
\[
E[Z_i Z_j] = P(Z_i = 1 \text{ and } Z_j = 1)
= P(Z_j = 1 \mid Z_i = 1)P(Z_i = 1)
= \left( \frac{n-1}{N-1} \right) \left( \frac{n}{N} \right).
\]

Because the population is finite, the $Z_i$’s are not quite independent—if we know that unit $i$ is in the sample, we do have a small amount of information about whether unit $j$ is
in the sample, reflected in the conditional probability \( P(Z_j = 1 \mid Z_i = 1) \). Consequently, for \( i \neq j \), the covariance of \( Z_i \) and \( Z_j \) is:

\[
\text{Cov}(Z_i, Z_j) = E[Z_i Z_j] - E[Z_i]E[Z_j]
\]

\[
= \frac{n - 1}{N - 1} - \left( \frac{n}{N} \right)^2
\]

\[
= -\frac{1}{N - 1} \left( 1 - \frac{n}{N} \right) \left( \frac{n}{N} \right).
\]

The negative covariance of \( Z_i \) and \( Z_j \) is the source of the fpc. The following derivation shows how we can use the random variables \( Z_1, \ldots, Z_N \) and the properties of covariances given in Appendix A to find \( V(\bar{y}) \):

\[
V(\bar{y}) = \frac{1}{n^2} V \left( \sum_{i=1}^{N} Z_i y_i \right)
\]

\[
= \frac{1}{n^2} \text{Cov} \left( \sum_{i=1}^{N} Z_i y_i, \sum_{j=1}^{N} Z_j y_j \right)
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \text{Cov}(Z_i, Z_j)
\]

\[
= \frac{1}{n^2} \left[ \sum_{i=1}^{N} y_i^2 V(Z_i) + \sum_{i=1}^{N} \sum_{j \neq i} y_i y_j \text{Cov}(Z_i, Z_j) \right]
\]

\[
= \frac{1}{n^2} \left[ \frac{n}{N} \left( 1 - \frac{n}{N} \right) \sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} \sum_{j \neq i} y_i y_j \frac{1}{N-1} \left( 1 - \frac{n}{N} \right) \left( \frac{n}{N} \right) \right]
\]

\[
= \frac{1}{n^2} \frac{n}{N} \left( 1 - \frac{n}{N} \right) \left[ \sum_{i=1}^{N} y_i^2 - \frac{1}{N-1} \sum_{i=1}^{N} \sum_{j \neq i} y_i y_j \right]
\]

\[
= \frac{1}{n} \left( 1 - \frac{n}{N} \right) \frac{1}{N(N-1)} \left[ (N-1) \sum_{i=1}^{N} y_i^2 - \left( \sum_{i=1}^{N} y_i \right)^2 + \sum_{i=1}^{N} y_i^2 \right]
\]

\[
= \frac{1}{n} \left( 1 - \frac{n}{N} \right) \frac{1}{N(N-1)} \left[ N \sum_{i=1}^{N} y_i^2 - \left( \sum_{i=1}^{N} y_i \right)^2 \right]
\]

\[
= \left( 1 - \frac{n}{N} \right) \frac{S^2}{n}.
\]

To show that the estimator in (2.11) is an unbiased estimator of the variance, we need to show that \( E[s^2] = S^2 \). The argument proceeds much like the previous one. Since \( S^2 = \sum_{i=1}^{N} (y_i - \bar{y})^2 / (N - 1) \), it makes sense when trying to find an unbiased estimator to find the expected value of \( \sum_{i \in S} (y_i - \bar{y})^2 \), and then find the multiplicative
constant that will give the unbiasedness:

\[
E \left[ \sum_{i \in S} (y_i - \bar{y})^2 \right] = E \left[ \sum_{i \in S} ((y_i - \bar{y}_U) - (\bar{y} - \bar{y}_U))^2 \right] \\
= E \left[ \sum_{i \in S} (y_i - \bar{y}_U)^2 - n (\bar{y} - \bar{y}_U)^2 \right] \\
= E \left[ \sum_{i=1}^{N} Z_i (y_i - \bar{y}_U)^2 \right] - n \text{V}(\bar{y}) \\
= \frac{n}{N} \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 - \left( 1 - \frac{n}{N} \right) S^2 \\
= \frac{n(N-1)}{N} S^2 - \frac{N-n}{N} S^2 \\
= (n-1)S^2.
\]

Thus,

\[
E \left[ \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y})^2 \right] = E[s^2] = S^2.
\]

### 2.9 A Prediction Approach for Simple Random Sampling*

Unless you have studied randomization theory in the design of experiments, the proofs in the preceding section probably seemed strange to you. The random variables in randomization theory are not concerned with the responses \(y_i\). The random variables \(Z_1, \ldots, Z_N\) are indicator variables that tell us whether the \(i\)th unit is in the sample or not. In a design-based, or randomization-theory, approach to sampling inference, the only relationship between units sampled and units not sampled is that the nonsampled units could have been sampled had we used a different starting value for the random number generator.

In Section 2.8 we found properties of the sample mean \(\bar{y}\) using randomization theory: \(y_1, y_2, \ldots, y_N\) were considered to be fixed values, and \(\bar{y}\) is unbiased because \(\bar{y} = (1/N) \sum_{i=1}^{N} Z_i y_i\) and \(E[Z_i] = P(Z_i = 1) = n/N\). The only probabilities used in finding the expected value and variance of \(\bar{y}\) are the probabilities that subsets of units are included in the sample. The quantity measured on unit \(i\), \(y_i\) can be anything: Whether \(y_i\) is number of television sets owned, systolic blood pressure, or acreage devoted to soybeans, the properties of estimators depend exclusively on the joint distribution of the random variables \(\{Z_1, \ldots, Z_N\}\).

In your other statistics classes, you most likely learned a different approach to inference, an approach explained in Chapter 5 of Casella and Berger (2002). There, you had random variables \(\{Y_i\}\) that followed some probability distribution, and the actual sample values were realizations of those random variables. Thus you assumed,
3. Stratified Sampling

3.1 Justification of stratified sampling

Stratified sampling is different from cluster sampling. In both cases the population is divided into subgroups: strata in the former and clusters in the latter. In cluster sampling only a portion of clusters are sampled while in stratified sampling every stratum will be sampled. Usually, only a subset of the elements in a stratum are observed, while all elements in a sampled cluster are observed.

Questions associated with stratified sampling include (i) Why use stratified sampling? (ii) How to stratify? and (iii) How to allocate sample sizes to each stratum? There are four main reasons to justify the use of stratified sampling:
• Administrative convenience. A survey at national level can be greatly facilitated if officials associated with each province survey a portion of the sample from their province. Here provinces are the natural choice of strata.

• In addition to the estimates for the entire population, estimates for certain sub-population are also required.

• Protect from possible disproportional samples under probability sampling.

• Increased accuracy of estimate. Stratified sampling can often provide more accurate estimates than SRS.
totals. Thus we estimate the total number of acres devoted to farming as 909,736,034, with standard error (SE) $\sqrt{2.5419 \times 10^{15}} = 50,417,248$. 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 population total in Example 2.5, 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 smaller standard error than an SRS with the same number of observations. The sample variance in Example 2.5 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 $7.647 \times 10^9$.

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

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

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.

### 3.2 Theory of Stratified Sampling

We divide the population of $N$ sampling units into $H$ “layers” or strata, with $N_h$ sampling units in stratum $h$. 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 $N_h$ population units in stratum $h$. Define $S_h$ to be the set of $n_h$ units in the SRS for stratum $h$. The total sample size is $n = n_1 + n_2 + \cdots + n_H$. 

Notation for Stratification: The population quantities are:

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

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

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

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

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

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

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

\[ \bar{y}_h = \frac{1}{n_h} \sum_{j \in S_h} y_{hj} \]

\[ \hat{t}_h = \frac{N_h}{n_h} \sum_{j \in S_h} y_{hj} = N_h \bar{y}_h \]

\[ s^2_h = \frac{\sum_{j \in S_h} (y_{hj} - \bar{y}_h)^2}{n_h - 1} \]

Suppose we only sampled the \( h \)th 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

\[ \hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h \bar{y}_h. \] (3.1)

To estimate \( \bar{y}_U \), then, we use

\[ \bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h. \] (3.2)

This is a weighted average of the sample stratum averages; \( \bar{y}_h \) is multiplied by \( N_h / N \), the proportion of the population units in stratum \( h \). 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.** \( \bar{y}_{str} \) and \( \hat{t}_{str} \) are unbiased estimators of \( \bar{y}_U \) and \( t \). An SRS is taken in each stratum, so (2.30) implies that
  \[
  E \left[ \bar{y}_h \right] = \bar{y}_{hU} \quad \text{and} \quad E \left[ \hat{y}_h \right] = \bar{y}_{hU} = \hat{y}_U.
  \]

- **Variance of the estimators.** Since we are sampling independently from the strata, and we know \( V(\hat{t}_h) \) from the SRS theory, the properties of expected value in Section A.2 and (2.16) imply that
  \[
  V(\hat{t}_{str}) = \sum_{h=1}^{H} V(\hat{t}_h) = \sum_{h=1}^{H} \left( 1 - \frac{n_h}{N_h} \right) N_h^2 \frac{s_h^2}{n_h}, \quad (3.3)
  \]

- **Standard errors for stratified samples.** We can obtain an unbiased estimator of \( V(\hat{t}_{str}) \) by substituting the sample estimators \( s_h^2 \) for the population parameters \( S_h^2 \). Note that in order to estimate the variances, we need to sample at least two units from each stratum.
  \[
  \hat{V}(\hat{t}_{str}) = \sum_{h=1}^{H} \left( 1 - \frac{n_h}{N_h} \right) N_h^2 \frac{s_h^2}{n_h}, \quad (3.4)
  \]
  \[
  \hat{V}(\bar{y}_{str}) = \frac{1}{N^2} \hat{V}(\hat{t}_{str}) = \sum_{h=1}^{H} \left( 1 - \frac{n_h}{N_h} \right) \left( \frac{N_h}{N} \right)^2 \frac{S_h^2}{n_h}. \quad (3.5)
  \]

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

- **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 population mean \( \bar{y}_U \) is
  \[
  \bar{y}_{str} \pm z_{\alpha/2} \text{SE}(\bar{y}_{str}).
  \]

The central limit theorem used for constructing this CI is stated in Krewski and Rao (1981). Some survey software packages use the percentile of a \( t \) distribution with \( n - H \) degrees of freedom (df) rather than the percentile of the normal distribution.

**Example 3.3** Siniff and Skoog (1964) used stratified random sampling to estimate the size of the Nelchina herd of Alaska caribou in February of 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
Table 3.1

Spreadsheet for Calculations in Example 3.3

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stratum</td>
<td>$N_h$</td>
<td>$n_h$</td>
<td>$\bar{y}_h$</td>
<td>$s^2_h$</td>
<td>$\hat{t}_h = \frac{N_h \bar{y}_h}{N_h}$</td>
<td>$(1 - \frac{n_h}{N_h})N_h \frac{s^2_h}{n_h}$</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>400</td>
<td>98</td>
<td>24.1</td>
<td>5,575</td>
<td>9,640</td>
<td>6,872,040.82</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>30</td>
<td>10</td>
<td>25.6</td>
<td>4,064</td>
<td>768</td>
<td>243,840.00</td>
</tr>
<tr>
<td>4</td>
<td>C</td>
<td>61</td>
<td>37</td>
<td>267.6</td>
<td>347,556</td>
<td>16,323.6</td>
<td>13,751,945.51</td>
</tr>
<tr>
<td>5</td>
<td>D</td>
<td>18</td>
<td>6</td>
<td>179.0</td>
<td>22,798</td>
<td>3,222</td>
<td>820,728.00</td>
</tr>
<tr>
<td>6</td>
<td>E</td>
<td>70</td>
<td>39</td>
<td>293.7</td>
<td>123,578</td>
<td>20,559</td>
<td>6,876,006.67</td>
</tr>
<tr>
<td>7</td>
<td>F</td>
<td>120</td>
<td>21</td>
<td>33.2</td>
<td>9,795</td>
<td>3,984</td>
<td>5,541,171.43</td>
</tr>
<tr>
<td>8</td>
<td>total</td>
<td>211</td>
<td></td>
<td>54,496.6</td>
<td>34,105,732.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>$\sqrt{\text{total}}$</td>
<td></td>
<td></td>
<td>5,840.01</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$N_1 = 400$ sampling units; $n_1 = 98$ of these were randomly selected to be in the survey. The following data were reported:

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

The spreadsheet shown in Table 3.1 displays the calculations for finding the stratified sampling estimates. The estimated total number of caribou is 54,497 with standard error 5,840. An approximate 95% CI for the total number of caribou is

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

Of course, this CI only reflects the uncertainty due to sampling error; if the field procedure for counting caribou tends to miss animals, then the entire CI 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 simply use the results in (3.1)–(3.5), with $\bar{y}_h = \hat{p}_h$ and $s^2_h = \frac{n_h}{n_h - 1} \hat{p}_h (1 - \hat{p}_h)$. Then,

$$\hat{p}_{\text{str}} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h$$  \hspace{1cm} (3.6)
3.2 Theory of Stratified Sampling

and

\[
\hat{V}(\hat{p}_{\text{str}}) = \sum_{h=1}^{H} \left( 1 - \frac{n_h}{N_h} \right) \left( \frac{N_h}{N} \right)^2 \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}.
\] (3.7)

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

\[
\hat{t}_{\text{str}} = \sum_{h=1}^{H} N_h \hat{p}_h,
\]

so 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 3.4**

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 3.2.

Ignoring the nonresponse for now (we’ll return to the nonresponse in Exercise 7 of 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 that are female. Here, let \( N_h \) be the membership figures, and let \( n_h \) be the number of valid surveys. Thus,

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

and

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

**Table 3.2**

Data from ACLS Survey

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Membership</th>
<th>Number Mailed</th>
<th>Valid Returns</th>
<th>Female Members (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature</td>
<td>9,100</td>
<td>915</td>
<td>636</td>
<td>38</td>
</tr>
<tr>
<td>Classics</td>
<td>1,950</td>
<td>633</td>
<td>451</td>
<td>27</td>
</tr>
<tr>
<td>Philosophy</td>
<td>5,500</td>
<td>658</td>
<td>481</td>
<td>18</td>
</tr>
<tr>
<td>History</td>
<td>10,850</td>
<td>855</td>
<td>611</td>
<td>19</td>
</tr>
<tr>
<td>Linguistics</td>
<td>2,100</td>
<td>667</td>
<td>493</td>
<td>36</td>
</tr>
<tr>
<td>Political Science</td>
<td>5,500</td>
<td>833</td>
<td>575</td>
<td>13</td>
</tr>
<tr>
<td>Sociology</td>
<td>9,000</td>
<td>824</td>
<td>588</td>
<td>26</td>
</tr>
<tr>
<td>Totals</td>
<td>44,000</td>
<td>5,385</td>
<td>3,835</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3: Stratified Sampling

The estimated total number of female members in the societies is \( \hat{t}_{str} = 44,000 \times (0.2465) = 10,847 \), with \( SE(\hat{t}_{str}) = 44,000 \times (0.0071) = 312 \). □

3.3 Sampling Weights in Stratified Random Sampling

We introduced the notion of sampling weight, \( w_i = 1/\pi_i \), in Section 2.4. For an SRS, the sampling weight for each observation is the same since all of the inclusion probabilities \( \pi_i \) are the same. In stratified sampling, however, we may have different inclusion probabilities in different strata so that the weights may be unequal for some stratified sampling designs.

The stratified sampling estimator \( \hat{t}_{str} \) can be expressed as a weighted sum of the individual sampling units: Using (3.1),

\[
\hat{t}_{str} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \sum_{j \in S_h} \frac{N_h}{n_h} y_{hj}.
\]

The estimator of the population total in stratified sampling may thus be written as

\[
\hat{t}_{str} = \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}, \quad (3.8)
\]

where the sampling weight for unit \( j \) of stratum \( h \) is \( w_{hj} = (N_h/n_h) \). The sampling weight can again be thought of as the number of units in the population represented by the sample member \( y_{hj} \). 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 another woman not selected to be in the sample, and each man represents himself and seven other men not in the sample. Note that the probability of including unit \( j \) of stratum \( h \) in the sample is \( \pi_{hj} = n_h/N_h \), the sampling fraction in stratum \( h \). Thus, as before, the sampling weight is simply the reciprocal of the inclusion probability:

\[
w_{hj} = \frac{1}{\pi_{hj}}. \quad (3.9)
\]

The sum of the sampling weights in stratified random sampling 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. In a stratified random sample, the population mean is thus estimated by

\[
\bar{y}_{str} = \frac{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj}}. \quad (3.10)
\]
Chapter 4

Ratio and Regression Estimation

Often in survey sampling, information on one (or more) covariate \(x\) (called auxiliary variable) is available prior to sampling. Sometimes this auxiliary information is complete, i.e. the value \(x_i\) is known for every element \(i\) in the population; sometimes only the population mean \(\bar{X} = \frac{1}{N-1} \sum_{i=1}^{N} x_i\) or total \(X = \sum_{i=1}^{N} x_i\) is known. When the auxiliary variable \(x\) is correlated with the study variable \(y\), this known auxiliary information can be useful for the new survey study.

Example 4.1 In family expenditure surveys, the values on \(x^{(1)}\): the number of people in the family and/or \(x^{(2)}\): the family income of previous year are known for every element in the population. The study variable(s) is on current year family expenditures such as expenses on clothing, food, furniture, etc.

Example 4.2 In agriculture surveys, a complete list of farms with the area (acreage) of each farm is available.

Example 4.3 Data from earlier census provides various population totals that can be used as auxiliary information for the planned surveys.

Auxiliary information can be used at the design stage. For instance, a stratified sampling scheme might be chosen where stratification is done by values of certain covariates such as sex, age and income levels. The \(pps\) sampling design (inclusion probability proportional to a size measure) is another sophisticated example.

In this chapter, we use auxiliary information explicitly at the estimation
stage by incorporating the known $\bar{X}$ or $X$ into the estimators through ratio and regression estimation. The resulting estimators will be more efficient than those discussed in previous chapters.

4.1 Ratio estimator

4.1.1 Ratio estimator under SRSWOR

Suppose $y_i$ is approximately proportional to $x_i$, i.e. $y_i \approx \beta x_i$ for $i = 1, 2, \ldots, N$. It follows that $\bar{Y} \approx \beta \bar{X}$. Let $R = \bar{Y}/\bar{X} = Y/X$ be the ratio of two population means or totals. Let $\bar{y}$ and $\bar{x}$ be the sample means under SRSWOR. It is natural to use $\hat{R} = \bar{y}/\bar{x}$ to estimate $R$. The ratio estimator for $\bar{Y}$ is defined as

$$\hat{Y}_R = \hat{R}\bar{X} = \frac{\bar{y}}{\bar{x}}\bar{X}.$$ 

One can expect that $\bar{X}/\bar{x}$ will be close to 1, so $\hat{Y}_R$ will be close to $\bar{y}$. Why is ratio estimator often used? The following results will provide an answer.

Result 4.1 Under simple random sampling without replacement,

(i) $\hat{Y}_R$ is approximately unbiased estimator for $\bar{Y}$.

(ii) The variance of $\hat{Y}_R$ can be approximated by

$$V(\hat{Y}_R) \approx (1 - \frac{n}{N})\frac{1}{nN-1}\sum_{i=1}^{N}(y_i - Rx_i)^2.$$ 

(iii) An approximately unbiased variance estimator is given by

$$v(\hat{Y}_R) = (1 - \frac{n}{N})\frac{1}{nn-1}\sum_{i\in s}(y_i - \hat{R}x_i)^2.$$ 

To see when the ratio estimator is better than the simple sample mean $\bar{y}$, let's make a comparison between the two variances. Note that

$$V(\bar{y}) = (1 - \frac{n}{N})\frac{1}{n}\sigma_Y^2,$$
4.1. RATIO ESTIMATOR

\[
V(\hat{Y}_R) \approx (1 - \frac{n}{N}) \frac{1}{n} \frac{1}{N-1} \sum_{i=1}^{N} [(y_i - \bar{Y}) - R(x_i - \bar{X})]^2
\]

\[
= (1 - \frac{n}{N}) \frac{1}{n} [S_Y^2 + R^2 S_X^2 - 2RS_{XY}],
\]

where \(S_Y^2\) and \(S_X^2\) are the population variances for the y and x variables, and \(S_{XY} = (N - 1)^{-1} \sum_{i=1}^{N} (y_i - \bar{Y})(x_i - \bar{X})\). The ratio estimator will have a smaller variance if and only if

\[
R^2 S_X^2 - 2RS_{XY} < 0.
\]

This condition can also be re-expressed as

\[
\rho > \frac{1}{2} \frac{CV(X)}{CV(Y)},
\]

where \(\rho = S_{XY}/[S_X S_Y]\), \(CV(X) = S_X/\bar{X}\) and \(CV(Y) = S_Y/\bar{Y}\). The conclusion is: if there is a strong correlation between y and x, the ratio estimator will perform better than the simple sample mean. Indeed, in many practical situations \(CV(X) = CV(Y)\), we only require \(\rho > 0.5\). This is usually the case.

A scatter plot of the data can visualize the relationship between y and x. If a straight line going through the origin is appropriate, ratio estimator may be efficient.

Ratio estimator can provide improved estimate. There are other situations where we have to use a ratio type estimator. Under one-stage cluster sampling with clusters of unequal sizes and \(M_i\) are not known unless the \(i\)th cluster is selected in the sample, the population mean (per element) is indeed a ratio:

\[
\bar{Y} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} / \sum_{i=1}^{N} M_i = \left[ \frac{1}{N} \sum_{i=1}^{N} Y_i / \left[ \frac{1}{N} \sum_{i=1}^{N} M_i \right] \right].
\]

A natural estimate for \(\bar{Y}\) would be \(\hat{\bar{Y}} = [n^{-1} \sum_{i \in s} Y_i] / [n^{-1} \sum_{i \in s} M_i]\).

4.1.2 Ratio estimator under stratified random sampling

When the population has been stratified, ratio estimator can be used in two different ways: (a) estimate \(R = \bar{Y}/\bar{X}\) by \(\hat{R} = \bar{y}_{st}/\bar{x}_{st}\), and \(\bar{Y} = R\bar{X}\) by \(\hat{R}\bar{X}\);
of students planning a career in teaching accordingly. To estimate the total number
of women who plan to follow a career in teaching, let
\[
  y_i = \begin{cases} 
  1 & \text{if woman and plans career in teaching} \\
  0 & \text{otherwise.}
  \end{cases}
\]
\[
  x_i = \begin{cases} 
  1 & \text{if woman} \\
  0 & \text{otherwise.}
  \end{cases}
\]

Then \((84/240) \times 2700 = (\bar{y}/\bar{x})t_x\) is a ratio estimate of the total number of women planning a career in teaching. Similarly, \((40/160) \times 1300\) is a ratio estimate of the total number of men planning a teaching career.

This use of ratio estimation, called poststratification, will be discussed in Section 4.4 and Chapters 7 and 8.

5 Ratio estimation may be used to adjust for nonresponse, as will be discussed in Chapter 8. Suppose a sample of businesses is taken; let \(y_i\) be the amount spent on health insurance by business \(i\) and \(x_i\) be the number of employees in business \(i\). Assume that \(t_x\), the total number of employees in all businesses in the population, is known. We expect that the amount a business spends on health insurance will be related to the number of employees. Some businesses may not respond to the survey, however. One method of adjusting for the nonresponse when estimating total insurance expenditures is to multiply the ratio \(\bar{y}/\bar{x}\) (using data only from the respondents) by the population total \(t_x\). If companies with few employees are less likely to respond to the survey, and if \(y_i\) is proportional to \(x_i\), then we would expect the estimate \(N\bar{y}\) to overestimate the population total \(t_x\). In the ratio estimate \(t_x\bar{y}/\bar{x}\), \(t_x/\bar{x}\) is likely to be smaller than \(N\) because companies with many employees are more likely to respond to the survey. Thus a ratio estimate of total health care insurance expenditures may help to compensate for the nonresponse of companies with few employees.

**Example 4.2** Let’s return to the SRS from the U.S. Census of Agriculture, described in Example 2.5. The file agsrs.dat contains data from an SRS of 300 of the 3078 counties.

For this example, suppose we know the population totals for 1987, but only have 1992 information on the SRS of 300 counties. When the same quantity is measured at different times, the response of interest at an earlier time often makes an excellent auxiliary variable. Let
\[
  y_i = \text{total acreage of farms in county } i \text{ in } 1992
\]
\[
  x_i = \text{total acreage of farms in county } i \text{ in } 1987.
\]
In 1987 a total of \(t_x = 964,470,625\) acres were devoted to farms in the United States. The average acreage per county for the population is then \(\bar{x}_U = 964,470,625/3078 = 313,343.3\) acres of farms per county. The data, and the line through the origin with slope \(\hat{B}\), are plotted in Figure 4.1.

A portion of a spreadsheet with the 300 values of \(x_i\) and \(y_i\) is given in Table 4.1. Cells C305 and D305 contain the sample averages of \(y\) and \(x\), respectively, so
\[
\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{C305}{D305} = 0.986565,
\]
Chapter 4: Ratio and Regression Estimation

FIGURE 4.1
The plot of acreage, 1992 vs. 1987, for an SRS of 300 counties. The line in the plot goes through the origin and has slope \( \hat{B} = 0.9866 \). Note that the variability about the line increases with \( x \).

\[
\hat{y}_r = \hat{B} \bar{x}_U = (\hat{B})(313,343.283) = 309,133.6,
\]

and

\[
\hat{t}_{yr} = \hat{B}t_x = (\hat{B})(964,470,625) = 951,513,191.
\]

Note that \( \bar{y} \) for these data is 297,897.0, so \( \hat{t}_{y_{SRS}} = (3078)(\bar{y}) = 916,927,110 \). In this example, \( \bar{x}_S = 301,953.7 \) is smaller than \( \bar{x}_U = 313,343.3 \). This means that our SRS of size 300 slightly underestimates the true population mean of the \( x \)'s. Since the \( x \)'s and \( y \)'s are positively correlated, we have reason to believe that \( \bar{y}_S \) may also underestimate the population value \( \bar{y}_U \). Ratio estimation gives a more precise estimate of \( \bar{y}_U \) by expanding \( \bar{y}_S \) by the factor \( \bar{x}_U / \bar{x}_S \). Figure 4.2 shows the ratio and SRS estimates of \( \bar{y}_U \) on a graph of the center part of the data.

4.1.2 Bias and Mean Squared Error of Ratio Estimators

Unlike the estimators \( \bar{y} \) and \( N\bar{y} \) in an SRS, ratio estimators are usually biased for estimating \( \bar{y}_U \) and \( t_y \). We start with the unbiased estimator \( \bar{y} \)—if we calculate \( \bar{y}_S \) for each possible SRS \( S \), then the average of all of the sample means from the possible samples is the population mean \( \bar{y}_U \). The estimation bias in ratio estimation arises
4.1 Ratio Estimation in a Simple Random Sample

### Table 4.1
Part of the Spreadsheet for the Census of Agriculture Data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 County</td>
<td>State</td>
<td>acres92 (y)</td>
<td>acres87 (x)</td>
<td>Residual (y − Bx)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Coffee County</td>
<td>AL</td>
<td>175209</td>
<td>179311</td>
<td>−1693.00</td>
</tr>
<tr>
<td>4 Colbert County</td>
<td>AL</td>
<td>138135</td>
<td>145104</td>
<td>−5019.56</td>
</tr>
<tr>
<td>5 Lamar County</td>
<td>AL</td>
<td>56102</td>
<td>59861</td>
<td>−2954.78</td>
</tr>
<tr>
<td>6 Marengo County</td>
<td>AL</td>
<td>199117</td>
<td>220526</td>
<td>−18446.29</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>299 Rock County</td>
<td>WI</td>
<td>343115</td>
<td>357751</td>
<td>−9829.70</td>
</tr>
<tr>
<td>300 Kanawha County</td>
<td>WV</td>
<td>19956</td>
<td>21369</td>
<td>−1125.91</td>
</tr>
<tr>
<td>301 Pleasants County</td>
<td>WV</td>
<td>15650</td>
<td>15716</td>
<td>145.14</td>
</tr>
<tr>
<td>302 Putnam County</td>
<td>WV</td>
<td>55827</td>
<td>55635</td>
<td>939.44</td>
</tr>
<tr>
<td>303</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>304 Column sum</td>
<td></td>
<td>89369114</td>
<td>90586117</td>
<td>3.96176E-09</td>
</tr>
<tr>
<td>305 Column average</td>
<td></td>
<td>297897.0467</td>
<td>301953.7233</td>
<td></td>
</tr>
<tr>
<td>306 Column standard deviation</td>
<td></td>
<td>344551.8948</td>
<td>344829.5964</td>
<td>31657.21817</td>
</tr>
<tr>
<td>307 $\hat{B} = C305/D305 =$</td>
<td></td>
<td>0.986565237</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Figure 4.2
Detail of the center portion of Figure 4.1. Here, $\bar{x}_U$ is larger than $\bar{x}_S$, so $\hat{y}_r$ is larger than $\hat{y}_S$. 
In large samples, the bias of \( \hat{\bar{y}}_r \) is typically small relative to \( V(\hat{\bar{y}}_r) \), so that \( \text{MSE}(\hat{\bar{y}}_r) \approx V(\hat{\bar{y}}_r) \) (see Exercise 21). Thus, in the following, we use \( \hat{V}(\hat{\bar{y}}_r) \) to estimate both the variance and the MSE.

Note from (4.8) that
\[
E\left[ (\bar{y} - B\bar{x})^2 \right] = V(\bar{d}), \quad \text{where} \quad d_i = y_i - Bx_i \quad \text{and} \quad \bar{d}_U = 0.
\]

Since the deviations \( d_i \) depend on the unknown value \( B \), define the new variable
\[
e_i = y_i - \hat{B}x_i,
\]
which is the \( i \)-th residual from fitting the line \( y = \hat{B}x \). Estimate \( V(\hat{\bar{y}}_r) \) by
\[
\hat{V}(\hat{\bar{y}}_r) = \left( 1 - \frac{n}{N} \right) \left( \frac{\bar{x}_U}{\bar{x}} \right)^2 \frac{s_e^2}{n}, \quad (4.9)
\]
where \( s_e^2 \) is the sample variance of the residuals \( e_i \):
\[
s_e^2 = \frac{1}{n-1} \sum_{i \in S} e_i^2.
\]
[Exercise 19 explains why we include the factor \( \bar{x}_U/\bar{x} \) in (4.9). In large samples, we expect \( \bar{x}_U/\bar{x} \) to be approximately equal to 1.] Similarly,
\[
\hat{V}(\hat{B}) = \left( 1 - \frac{n}{N} \right) \frac{s_e^2}{n\bar{x}^2}, \quad (4.10)
\]
and
\[
\hat{V}(\hat{t}_y) = \hat{V}(t_x\hat{B}) = \left( 1 - \frac{n}{N} \right) \left( \frac{t_x}{\bar{x}} \right)^2 \frac{s_e^2}{n}, \quad (4.11)
\]
If the sample sizes are sufficiently large, approximate 95% confidence intervals (CIs) can be constructed using the standard errors (SEs) as
\[
\hat{B} \pm 1.96 \text{SE}(\hat{B}), \quad \hat{\bar{y}}_r \pm 1.96 \text{SE}(\hat{\bar{y}}_r), \quad \text{or} \quad \hat{t}_y \pm 1.96 \text{SE}(\hat{t}_y).
\]

Some software packages, including SAS software, substitute a \( t \) percentile with \( n - 1 \) degrees of freedom for the normal percentile 1.96.

**Example 4.3** Let’s return to the sample taken from the Census of Agriculture in Example 4.2. In the spreadsheet in Table 4.1, we created Column E, containing the residuals \( e_i = y_i - \hat{B}x_i \). The sample standard deviation of Column E was calculated in Cell E306 to be \( s_e = 31,657.218 \). Thus, using (4.11),
\[
\text{SE}(\hat{t}_y) = 3078\sqrt{1 - \frac{300}{3078}} \left( \frac{313,343.283}{301,953.723} \right) \frac{31,657.218}{\sqrt{300}} = 5,546,162.
\]
An approximate 95% CI for the total farm acreage, using the ratio estimator, is
\[
951,513,191 \pm 1.96(5,546,162) = [940,642,713, \ 962,383,669].
\]
The website gives SAS code for calculating the ratio $\hat{B} = \bar{y}/\bar{x}$ and its standard error, with the output:

\[
\begin{array}{lrrrr}
\text{Ratio Analysis: acres92/acres87} \\
\hline
\text{Numerator} & \text{Denominator} & \text{Ratio} & \text{Std Err} & 95\% \text{ CL for Ratio} \\
\hline
\text{acres92} & \text{acres87} & 0.986565 & 0.005750 & 0.97524871 & 0.99788176 \\
\hline
\end{array}
\]

We then multiply each quantity in the output by $t_x = 964,470,625$ (we do the calculations on the computer to avoid roundoff error) to obtain $\hat{t}_y = (964,470,625)(0.986565237) = 951,131,191$ and 95% CI for $t_y$ of $\left[940,598,734, 962,427,648\right]$. SAS PROC SURVEYMEANS uses the percentile from a $t_{299}$ distribution, 1.96793, instead of the value 1.96 from the normal distribution, so the CI from SAS software is slightly larger than the one we obtained when doing calculations by hand.

Did using a ratio estimator for the population total improve the precision in this example? The standard error of $\hat{t}_y = N\bar{y}$ is more than 10 times as large:

\[
\text{SE}(N\bar{y}) = 3078 \sqrt{\left(1 - \frac{150}{3078}\right) \frac{s_y}{\sqrt{150}}} = 58,169,381.
\]

The estimated CV for the ratio estimator is 5,546,162/951,131,191 = 0.0058, as compared with an estimated CV of 0.0634 for the SRS estimator $N\bar{y}$ which does not use the auxiliary information. Including the 1987 information through the ratio estimator has greatly increased the precision. If all quantities to be estimated were highly correlated with the 1987 acreage, we could dramatically reduce the sample size and still obtain high precision by using ratio estimators rather than $N\bar{y}$. 

\[\blacksquare\]

**Example 4.4** Let’s take another look at the hypothetical population used in Example 2.2 and in Exercise 2 of Chapter 3. Now, though, instead of using $x$ as a stratification variable in stratified sampling, we use it as auxiliary information for ratio estimation. The population values are the following:

<table>
<thead>
<tr>
<th>Unit Number</th>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>
Thus, the ratio estimator is calibrated to the population total $t_x$ of the $x$ variable, but is no longer calibrated to the population size $N$.

### 4.1.5 Advantages of Ratio Estimation

Ratio estimation is motivated by the desire to use information about a known auxiliary quantity $x$ to obtain a more accurate estimator of $t_y$ or $\bar{y}$. If $x$ and $y$ are perfectly correlated, that is, $y_i = Bx_i$ for all $i = 1, \ldots, N$, then $\hat{t}_y = t_y$ and there is no estimation error. In general, if $y_i$ is roughly proportional to $x_i$, the MSE will be small.

When does ratio estimation help? If the deviations of $y_i$ from $\hat{B}x_i$ are smaller than the deviations of $y_i$ from $\bar{y}$, then $\hat{V}(\hat{\bar{y}}_r) \leq \hat{V}(\bar{y})$. Recall from Chapter 2 that

$$ \text{MSE}(\bar{y}) = V(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}. $$

Using the approximation in (4.7) and (4.8),

$$ \text{MSE}(\hat{\bar{y}}_r) \approx \left(1 - \frac{n}{N}\right)\frac{1}{n}(S_y^2 - 2BR_xS_y + B^2S_x^2). $$

Thus,

$$ \text{MSE}(\hat{\bar{y}}_r) - \text{MSE}(\bar{y}) \approx \left(1 - \frac{n}{N}\right)\frac{1}{n}(S_y^2 - 2BR_xS_y + B^2S_x^2 - S_y^2) $$

$$ = \left(1 - \frac{n}{N}\right)\frac{1}{n}S_xB(-2RS_y + BS_x) $$

so to the accuracy of the approximation,

$$ \text{MSE}(\hat{\bar{y}}_r) \leq \text{MSE}(\bar{y}) \text{ if and only if } R \geq \frac{BS_x}{2S_y} = \frac{\text{CV}(x)}{2\text{CV}(y)}. $$

If the CVs are approximately equal, then it pays to use ratio estimation when the correlation between $x$ and $y$ is larger than 1/2.

Ratio estimation is most appropriate if a straight line through the origin summarizes the relationship between $x_i$ and $y_i$ and if the variance of $y_i$ about the line is proportional to $x_i$. Under these conditions, $\hat{B}$ is the weighted least squares regression slope for the line through the origin with weights proportional to $1/x_i$—the slope $\hat{B}$ minimizes the sum of squares

$$ \sum_{i \in S} \frac{1}{x_i}(y_i - \hat{B}x_i)^2. $$

### 4.2 Estimation in Domains

Often we want separate estimates for subpopulations; the subpopulations are called domains or subdomains. We may want to take an SRS of 1000 people from a population of 50,000 people and estimate the average salary for men and the average salary for women. There are two domains: men and women. We do not know which persons in the population belong to which domain until they are sampled, though. Thus, the
number of persons in an SRS who fall into each domain is a random variable, with value unknown at the time the survey is designed.

Estimating domain means is a special case of ratio estimation. Suppose there are $D$ domains. Let $\mathcal{U}_d$ be the index set of the units in the population that are in domain $d$, and let $\mathcal{S}_d$ be the index set of the units in the sample that are in domain $d$, for $d = 1, 2, \ldots, D$. Let $N_d$ be the number of population units in $\mathcal{U}_d$, and $n_d$ be the number of sample units in $\mathcal{S}_d$. Suppose we want to estimate the mean salary for the domain of women,

$$\bar{y}_{U_d} = \frac{\sum_{i \in \mathcal{U}_d} y_i}{N_d} = \frac{\text{total salary for all women in population}}{\text{number of women in population}}.$$

A natural estimator of $\bar{y}_{U_d}$ is

$$\tilde{y}_d = \frac{\sum_{i \in \mathcal{S}_d} y_i}{n_d} = \frac{\text{total salary for women in sample}}{\text{number of women in sample}},$$

which looks at first just like the sample means studied in Chapter 2.

The quantity $n_d$ is a random variable, however: If a different SRS is taken, we will very likely have a different value for $n_d$, the number of women in the sample. To see that $\tilde{y}_d$ uses ratio estimation, let

$$x_i = \begin{cases} 1 & \text{if } i \in \mathcal{U}_d \\ 0 & \text{if } i \notin \mathcal{U}_d, \end{cases}$$

$$u_i = y_i x_i = \begin{cases} y_i & \text{if } i \in \mathcal{U}_d \\ 0 & \text{if } i \notin \mathcal{U}_d. \end{cases}$$

Then $t_x = \sum_{i=1}^{N} x_i = N_d$, $\bar{x}_U = N_d / N$, $t_u = \sum_{i=1}^{N} u_i$, $\bar{y}_{U_d} = t_u / t_x = B$, $\bar{x} = n_d / n$, and

$$\bar{y}_d = \hat{B} = \frac{\bar{u}}{\bar{x}} = \frac{t_u}{t_x}.$$

Because we are estimating a ratio, we use (4.10) to calculate the standard error:

$$\text{SE}(\bar{y}_d) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{1}{n\bar{x}^2} \frac{\sum_{i \in \mathcal{S}_d} (u_i - \hat{B} x_i)^2}{n - 1}}$$

$$= \sqrt{\left(1 - \frac{n}{N}\right) \frac{1}{n\bar{x}^2} \frac{\sum_{i \in \mathcal{S}_d} (y_i - \hat{B})^2}{n - 1}}$$

$$= \sqrt{\left(1 - \frac{n}{N}\right) \frac{n}{n_d^2} \frac{(n_d - 1)s_{yd}^2}{n - 1}},$$

(4.13)

where

$$s_{yd}^2 = \frac{\sum_{i \in \mathcal{S}_d} (y_i - \bar{y}_d)^2}{n_d - 1}.$$
is the sample variance of the sample observations in domain \( d \). If \( E(n_d) \) is large, then \((n_d - 1)/n_d \approx 1\) and

\[
\text{SE}(\bar{y}_d) \approx \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_{yd}^2}{n_d}}.
\]

In a sufficiently large sample, the standard error of \( \bar{y}_d \) is approximately the same as if we used formula (2.12).

The situation is a little more complicated when estimating a domain total. If \( N_d \) is known, we can estimate \( t_u \) by \( N_d \bar{y}_d \). If \( N_d \) is unknown, though, we need to estimate \( t_u \) by

\[
\hat{t}_{yd} = \hat{t}_u = N\bar{u}.
\]

The standard error is

\[
\text{SE}(\hat{t}_{yd}) = N \text{SE}(\bar{u}) = N \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_u^2}{n}}.
\]

**Example 4.7**

In the SRS of size 300 from the Census of Agriculture (see Example 2.5), 39 counties are in western states.\(^2\) What is the estimated total number of acres devoted to farming in the West?

The sample mean of the 39 counties is \( \bar{y}_d = 598,681 \), with sample standard deviation \( s_{yd} = 516,157 \). Using (4.13),

\[
\text{SE}(\bar{y}_d) = \sqrt{\left(1 - \frac{300}{3078}\right) \frac{300}{39} \frac{38}{299} \frac{516,157.7}{\sqrt{39}}} = 77,637.
\]

Thus, \( \hat{CV}[\bar{y}_d] = 0.1297 \), and an approximate 95% CI for the mean farm acreage for counties in the western United States is [445,897, 751,463].

For estimating the total number of acres devoted to farming in the West, suppose we do not know how many counties in the population are in the western United States. Define

\[
x_i = \begin{cases} 
1, & \text{if county } i \text{ is in the western United States} \\
0, & \text{otherwise}
\end{cases}
\]

and define \( u_i = y_i x_i \). Then

\[
\hat{t}_{yd} = \hat{t}_u = \sum_{i \in S} \frac{3078}{300} u_i = 239,556,051. \quad (4.14)
\]

The standard error is

\[
\text{SE}(\hat{t}_{yd}) = 3078 \sqrt{\left(1 - \frac{300}{3078}\right) \frac{273005.4}{\sqrt{300}}} = 46,090,460.
\]

The estimated CV for \( \hat{t}_{yd} \) is \( \hat{CV}[\hat{t}_{yd}] = 46,090,460/239,556,051 = 0.1924 \); had we known the number of counties in the western United States and been able to use that

\(^2\)Alaska (AK), Arizona (AZ), California (CA), Colorado (CO), Hawaii (HI), Idaho (ID), Montana (MT), Nevada (NV), New Mexico (NM), Oregon (OR), Utah (UT), Washington (WA), and Wyoming (WY).
value in the estimate, the CV for the estimated total would have been 0.1297, the CV for the mean.

The SAS program on the website also contains the code for finding domain estimates. We define the domain indicator \( \text{west} \) to be 1 if the county is in the West and 0 otherwise. The relevant output is

\[
\begin{array}{cccccc}
\text{Domain Analysis: } \text{west} \\
\text{west} & \text{Variable} & \text{Mean} & \text{Std Error} & 95\% \text{ CL for Mean} \\
0 & \text{acres92} & 252952 & 16834 & 219825.176 & 286079.583 \\
1 & \text{acres92} & 598681 & 77637 & 445897.252 & 751463.927 \\
\end{array}
\]

\[
\begin{array}{cccccc}
\text{Domain Analysis: } \text{west} \\
\text{west} & \text{Variable} & \text{Sum} & \text{Std Dev} & 95\% \text{ CL for Sum} \\
0 & \text{acres92} & 677371058 & 47317687 & 584253179 & 770488938 \\
1 & \text{acres92} & 239556051 & 46090457 & 148853274 & 330258829 \\
\end{array}
\]

The output gives the estimates and CIs for both domains.

**Example 4.8** An SRS of 1500 licensed boat owners in a state was sampled from a list of 400,000 names with currently licensed boats; 472 of the respondents said they owned an open motorboat longer than 16 feet. The 472 respondents with large motorboats reported having the following numbers of children:

<table>
<thead>
<tr>
<th>Number of Children</th>
<th>Number of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>1</td>
<td>139</td>
</tr>
<tr>
<td>2</td>
<td>166</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>472</strong></td>
</tr>
</tbody>
</table>

If we are interested in characteristics of persons who own large motorboats, there are two domains: persons who own large motorboats (domain 1) and persons who do not own large motorboats (domain 2). To estimate the percentage of large-motorboat owners who have children, we can use \( \hat{p}_1 = 396/472 = 0.839 \). This is a ratio estimator, but in this case, as shown in (4.13), the standard error is approximately
what you would think it would be. Ignoring the fpc,
\[ SE(\hat{p}_1) = \sqrt{\frac{0.839(1 - 0.839)}{472}} = 0.017. \]

To look at the average number of children per household among registered boat owners who register a motorboat more than 16 feet long, note that the average number of children for the 472 respondents in the domain is 1.667373, with variance 1.398678. Thus an approximate 95% CI for the average number of children in large-motorboat households is
\[ 1.667 \pm 1.96\sqrt{\frac{1.398678}{472}} = [1.56, 1.77]. \]

To estimate the total number of children in the state whose parents register a large motorboat, we create a new variable \( u \) for the respondents that equals the number of children if the respondent has a motorboat, and 0 otherwise. The frequency distribution for the variable \( u \) is then

<table>
<thead>
<tr>
<th>Number of Children</th>
<th>Number of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1104</td>
</tr>
<tr>
<td>1</td>
<td>139</td>
</tr>
<tr>
<td>2</td>
<td>166</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>1500</td>
</tr>
</tbody>
</table>

Now \( \bar{u} = 0.52466 \) and \( s_u^2 = 1.0394178 \), so \( \hat{t}_{yd} = 400,000(0.524666) = 209,867 \) and
\[ SE(\hat{t}_{yd}) = SE(\hat{t}_u) = \sqrt{\left(1 - \frac{1500}{400,000}\right) \left(\frac{400,000}{400,000}\right)^2 \frac{1.0394178}{1500}} = 10,510. \]

The variable \( u_i \) counts the number of children in household \( i \) that belong to a household with a large open motorboat. SAS code to find estimates for this example is given on the website.

In this section, we have shown that estimating domain means is a special case of ratio estimation because the sample size in the domain varies from sample to sample. If the sample size for the domain in an SRS is sufficiently large, we can use SRS formulas for inference about the domain mean.

Inference about totals depends on whether the population size of the domain, \( N_d \), is known. If \( N_d \) is known, then the estimated total is \( N_d \bar{y}_d \). If \( N_d \) is unknown, then define a new variable \( u_i \) that equals \( y_i \) for observations in the domain and 0 for observations not in the domain; then use \( \hat{t}_u \) to estimate the domain total.
or (b) write $\bar{Y}$ as $\bar{Y} = \sum_{h=1}^{H} W_h \bar{Y}_h$ and estimate $\bar{Y}_h$, the strata mean, by a ratio estimator $[\bar{y}_h/\bar{x}_h]\bar{X}_h$. In (a), only $\bar{X}$ needs be known; under (b), the stratum means $\bar{X}_h$ are required.

The combined ratio estimator of $\bar{Y}$ is defined as

$$\hat{Y}_{Rc} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X},$$

where $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$, $\bar{x}_{st} = \sum_{h=1}^{H} W_h \bar{x}_h$. The separate ratio estimator of $\bar{Y}$ is defined as

$$\hat{Y}_{Rs} = \sum_{h=1}^{H} W_h \bar{y}_h / \bar{x}_h \bar{X}_h ,$$

where the $\bar{X}_h$'s are the known strata means.

**Result 4.2** Under stratified random sampling, the combined ratio estimator $\hat{Y}_{Rc}$ is approximately unbiased for $\bar{Y}$, and its variance is given by

$$V(\hat{Y}_{Rc}) \doteq \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \left( \frac{1}{n_h N_h} \right)^2 \sum_{j=1}^{N_h} [(y_{hj} - \bar{Y}_h) - R(x_{hj} - \bar{X}_h)]^2,$$

which can be estimated by

$$v(\hat{Y}_{Rc}) \doteq \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \left( \frac{1}{n_h N_h} \right)^2 \sum_{j \in s_h} [(y_{hj} - \bar{y}_h) - \hat{R}(x_{hj} - \bar{x}_h)]^2,$$

where $R = \bar{Y}/\bar{X}$ and $\hat{R} = \bar{y}_{st}/\bar{x}_{st}$. ♦

**Result 4.3** Under stratified random sampling, the separate ratio estimator $\hat{Y}_{Rs}$ is approximately unbiased for $\bar{Y}$, and its variance is given by

$$V(\hat{Y}_{Rs}) \doteq \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \left( \frac{1}{n_h N_h} \right)^2 \sum_{j=1}^{N_h} (y_{hj} - \bar{R}_h x_{hj})^2,$$

which can be estimated by

$$v(\hat{Y}_{Rs}) \doteq \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \left( \frac{1}{n_h n_h} \right)^2 \sum_{j \in s_h} (y_{hj} - \hat{R}_h x_{hj})^2,$$

where $R_h = \bar{Y}_h/\bar{X}_h$ and $\hat{R}_h = \bar{y}_h/\bar{x}_h$. ♦
uses ratio estimation separately in each stratum, with

\[
\hat{V}(\hat{y}_{xy}) = \sum_{h=1}^{H} \hat{V}(\hat{t}_{yhr}).
\]

It can improve efficiency if the \(\hat{t}_{yhr} / \hat{t}_{xh}\) vary from stratum to stratum, but should not be used when strata sample sizes are small because each ratio is biased, and the bias can propagate through the strata. Note that poststratification (Section 4.4) is a special case of the separate ratio estimator.

The combined estimator has less bias when the sample sizes in some of the strata are small. When the ratios vary greatly from stratum to stratum, however, the combined estimator does not take advantage of the extra efficiency afforded by stratification as does the separate ratio estimator. Many survey software packages, including SAS, calculate the combined ratio estimator by default.

**Example 4.10** Steffey et al. (2006) describe the use of combined ratio estimation in the legal case *Labor Ready v. Gates McDonald*. The plaintiff alleged that the defendant had not thoroughly investigated claims for worker’s compensation, resulting in overpayments for these claims by the plaintiff. A total of \(N = 940\) claims were considered in 1997. For each of these, the incurred cost of the claim \((x_i)\) was known, and consequently the total amount of incurred costs was known to be \(t_x = \$9.407\) million. But the plaintiff contended that the incurred value amounts were unjustified, and that the assessed value \((y_i)\) of some claims after a thorough review would differ from the incurred value.

A sampling plan was devised for estimating the total assessed value of all 940 claims. Since it was expected that the assessed value would be highly correlated with the incurred costs, ratio estimation is desirable here. Two strata were sampled: Stratum 1 consisted of the claims in which the incurred cost exceeded \(\$25,000\), and stratum 2 consisted of the smaller claims (incurred cost less than \(\$25,000\)). Summary statistics for the strata are given in the following table, with \(r_h\) the sample correlation in stratum \(h\):

<table>
<thead>
<tr>
<th>Stratum</th>
<th>(N_h)</th>
<th>(n_h)</th>
<th>(\bar{x}_h)</th>
<th>(s_{xh})</th>
<th>(\bar{y}_h)</th>
<th>(s_{yh})</th>
<th>(r_h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>102</td>
<td>70</td>
<td>$59,549.55</td>
<td>$64,047.95</td>
<td>$38,247.80</td>
<td>$32,470.78</td>
<td>0.62</td>
</tr>
<tr>
<td>2</td>
<td>838</td>
<td>101</td>
<td>$5,718.84</td>
<td>$5,982.34</td>
<td>$3,833.16</td>
<td>$5,169.72</td>
<td>0.77</td>
</tr>
</tbody>
</table>

The sampling fraction was set much higher in stratum 1 than in stratum 2 because the variability is much higher in stratum 1 (the investigators used a modified form of the optimal allocation described in Section 3.4.2). We estimate

\[
\hat{t}_{x,\text{str}} = \sum_{h=1}^{2} \hat{t}_{xh} = (102)(59,549.55) + (838)(5,718.84) = 10,866,442.02
\]

\[
\hat{t}_{y,\text{str}} = \sum_{h=1}^{2} \hat{t}_{yh} = (102)(38,247.80) + (838)(3,833.16) = 7,113,463.68
\]
Chapter 4: Ratio and Regression Estimation

\[
\hat{B} = \frac{\hat{y}_{\text{str}}}{\hat{t}_{\text{str}}} = \frac{7,113,463.68}{10,866,442.02} = 0.654626755.
\]

Using formulas for variances of stratified samples,

\[
\hat{V}(\hat{t}_{\text{str}}) = \left(1 - \frac{70}{102}\right) (102)^2 \frac{(64,047.95)^2}{70} + \left(1 - \frac{101}{838}\right) (838)^2 \frac{(5982.34)^2}{101} = 410,119,750,555,
\]

\[
\hat{V}(\hat{t}_{\text{reg}}) = \left(1 - \frac{70}{102}\right) (102)^2 \frac{(32,470.78)^2}{70} + \left(1 - \frac{101}{838}\right) (838)^2 \frac{(5169.72)^2}{101} = 212,590,045,044,
\]

and

\[
\hat{Cov}(\hat{t}_{\text{str}}, \hat{t}_{\text{reg}}) = \left(1 - \frac{70}{102}\right) (102)^2 \frac{(32,470.78)(64,047.95)(0.62)}{70} + \left(1 - \frac{101}{838}\right) (838)^2 \frac{(5169.72)(5982.34)(0.77)}{101} = 205,742,464,829.
\]

Using the combined ratio estimator, the total assessed value of the claims is estimated by

\[
\hat{y}_{\text{rc}} = (9.407 \times 10^6)(0.654626755) = $6.158 \text{ million}
\]

with standard error

\[
\text{SE}(\hat{y}_{\text{rc}}) = \frac{10.866}{9.407} \sqrt{[2.126 + (0.6546)^2(4.101) - 2(0.6546)(2.057)] \times 10^{11}}
\]

\[
= $0.371 \text{ million}.
\]

We use \(169 = (\text{number of observations}) - (\text{number of strata})\) degrees of freedom for the CI. An approximate 95% CI for the total assessed value of the claims is \(6.158 \pm 1.97(0.371), \text{ or between $5.43 and $6.89 million.}\) Note that the CI for \(t_y\) does not contain the total incurred value (\(t_x\)) of $9.407 million. This supported the plaintiff’s case that the total incurred value was too high.

4.6 Model-Based Theory for Ratio and Regression Estimation

In the design-based theory presented in Sections 4.1 and 4.3, the form of the estimators \(\hat{y}_r\) and \(\hat{y}_{\text{reg}}\) were motivated by regression models. Properties of the estimators, however, depend only on the sampling design. Thus, we found in (4.8) that

\[
V(\hat{y}_r) \approx \frac{1}{n} \left(1 - \frac{n}{N}\right) \sum_{i=1}^{N} \frac{(y_i - Bx_i)^2}{N - 1}
\]

for an SRS. This variance approximation is derived from the simple random sampling formulas in Chapter 2 and does not rely
One of the questions that needs to be addressed is how to make a choice between $\hat{Y}_{Rc}$ and $\hat{Y}_{Rs}$. First, it depends on what kind of auxiliary information is available. The separate ratio estimator requires the strata means $\bar{X}_h$ being known. If only $\bar{X}$ is known, the combined ratio estimator will have to be used. Second, in terms of efficiency, the variance of $\hat{Y}_{Rc}$ depends on the “residuals” $e_{hj} = (y_{hj} - \bar{Y}_h) - R(x_{hj} - \bar{X}_h)$, which is equivalent to fit a single straight line across all the strata with a common slope; while for the separate ratio estimator this slope can be different for different strata. So in many situations $\hat{Y}_{Rs}$ will perform better than $\hat{Y}_{Rc}$. Third, the variance formula for the separate ratio estimator depends on the approximation to $\bar{y}_h/\bar{x}_h$. If the sample sizes within each stratum, $n_h$, are too small, the bias from using $\hat{Y}_{Rs}$ can be large. The bias from using $\hat{Y}_{Rc}$, however, will be smaller since the approximation is made to $\bar{y}_{st}/\bar{x}_{st}$, and the pooled sample size $n$ will usually be large.

### 4.2 Regression estimator

The study variable $y$ is often linearly related to the auxiliary variable $x$, i.e. $y_i = \beta_0 + \beta_1 x_i, i = 1, 2, \cdots, N$. So roughly we have $\bar{Y} = \beta_0 + \beta_1 \bar{X}$ and $\bar{y} = \beta_0 + \beta_1 \bar{x}$. This leads to the regression type estimator of $\bar{Y}$: $\hat{Y} = \bar{y} + \hat{\beta}_1(\bar{X} - \bar{x})$.  The $\hat{\beta}_1$ is usually unknown and is estimated by the least square estimator $\hat{\beta}_1$ from the sample data. More formally, under SRSWOR, the regression estimator of $\bar{Y}$ is defined as

$$\hat{Y}_{REG} = \bar{y} + \hat{B}(\bar{X} - \bar{x}),$$

where $\hat{B} = \sum_{i \in s} (y_i - \bar{y})(x_i - \bar{x})/\sum_{i \in s}(x_i - \bar{x})^2$.

**Result 4.4** Under SRSWOR, the regression estimator $\hat{Y}_{REG}$ is approximately unbiased for $\bar{Y}$. Its approximate variance is given by

$$V(\hat{Y}_{REG}) \approx (1 - n/N)(1/N) \sum_{i=1}^{N} e_i^2,$$

where $e_i = y_i - B_0 - Bx_i$, $B = \sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})/\sum_{i=1}^{N}(x_i - \bar{X})^2$, and $B_0 = \bar{Y} - B\bar{X}$. This variance can be estimated by

$$v(\hat{Y}_{REG}) = (1 - n/N)(1/n - 1/n) \sum_{i \in s} e_i^2,$$
where $\hat{e}_i = y_i - \hat{B}_0 - \hat{B}x_i$, $\hat{B} = \sum_{i \in s} (y_i - \bar{y})(x_i - \bar{x})/\sum_{i \in s}(x_i - \bar{x})^2$, and $\hat{B}_0 = \bar{y} - \hat{B}\bar{x}$.

It can be shown that

$$V(\hat{\bar{Y}}_{REG}) = \left(1 - \frac{n}{N}\right)\frac{1}{n}S_Y^2(1 - \rho^2),$$

where $\rho = S_{XY}/[S_X S_Y]$ is the population correlation coefficient between $y$ and $x$. Since $|\rho| \leq 1$, we have $V(\hat{\bar{Y}}_{REG}) \leq V(\bar{y})$ under SRSWOR. When $n$ is large, the regression estimator is always more efficient than the simple sample mean $\bar{y}$.

It can also be shown that $V(\hat{\bar{Y}}_{REG}) \leq V(\hat{\bar{Y}}_R)$. So regression estimator is preferred in most situations. Ratio estimators are still being used by many survey practitioners due to its simplicity. If a scatter plot of the data shows that a straight line going through the origin fits the data well, then the regression estimator and the ratio estimator will perform similarly. Both requires only $\bar{X}$ be known to compute the estimates under SRSWOR. Under stratified sampling, a combined regression estimator and a separate regression estimator can be developed similarly.
The results of this section are only for SRSs, and the approximations depend on having a sufficiently large sample so that $E(n_d)$ is large. In Section 14.2, we discuss estimating domain means and totals if the data are collected using other sampling designs, or when the domain sample sizes are small.

4.3 Regression Estimation in Simple Random Sampling

4.3.1 Using a Straight-Line Model

Ratio estimation works best if the data are well fit by a straight line through the origin. Sometimes, data appear to be evenly scattered about a straight line that does not go through the origin—that is, the data look as though the usual straight-line regression model

$$y = B_0 + B_1x$$

would provide a good fit.

Let $\hat{B}_1$ and $\hat{B}_0$ be the ordinary least squares regression coefficients of the slope and intercept. For the straight line regression model,

$$\hat{B}_1 = \frac{\sum_{i \in S} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in S} (x_i - \bar{x})^2} = \frac{r_{xy}}{s_x},$$

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x},$$

and $r$ is the sample correlation coefficient of $x$ and $y$.

In regression estimation, like ratio estimation, we use the correlation between $x$ and $y$ to obtain an estimator for $\bar{y}_U$ with (we hope) increased precision. Suppose we know $\bar{x}_U$, the population mean for the $x$’s. Then the regression estimator of $\bar{y}_U$ is the predicted value of $y$ from the fitted regression equation when $x = \bar{x}_U$:

$$\hat{y}_{reg} = \hat{B}_0 + \hat{B}_1 \bar{x}_U = \bar{y} + \hat{B}_1 (\bar{x}_U - \bar{x}).$$

(4.15)

If $\bar{x}$ from the sample is smaller than the population mean $\bar{x}_U$ and $x$ and $y$ are positively correlated, then we would expect $\hat{y}$ to also be smaller than $\bar{y}_U$. The regression estimator adjusts $\hat{y}$ by the quantity $\hat{B}_1 (\bar{x}_U - \bar{x})$.

Like the ratio estimator, the regression estimator is biased. Let $B_1$ be the least squares regression slope calculated from all the data in the population,

$$B_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x}_U)(y_i - \bar{y}_U)}{\sum_{i=1}^{N} (x_i - \bar{x}_U)^2} = \frac{RS_y}{S_x}.$$
Then, using (4.15), the bias of \( \hat{\bar{y}}_{\text{reg}} \) is given by

\[
E[\hat{\bar{y}}_{\text{reg}} - \bar{y}_U] = E[\bar{y} - \bar{y}_U] + E[\hat{B}_1(\bar{x}_U - \bar{x})] = -\text{Cov}(\hat{B}_1, \bar{x}).
\]  (4.16)

If the regression line goes through all of the points \((x_i, y_i)\) in the population, then the bias is zero: in that situation, \(\hat{B}_1 = B_1\) for every sample, so \(\text{Cov}(\hat{B}_1, \bar{x}) = 0\). As with ratio estimation, for large SRSs the MSE for regression estimation is approximately equal to the variance (see Exercise 29); the bias is often negligible in large samples.

The method used in approximating the MSE in ratio estimation can also be applied to regression estimation. Let \(d_i = y_i - (\bar{y}_U + \hat{B}_1(x_i - \bar{x}_U))\). Then,

\[
\text{MSE}(\hat{\bar{y}}_{\text{reg}}) = E[(\bar{y} - \hat{B}_1(\bar{x}_U - \bar{x}))^2] 
\approx \text{Var}(d) 
= \left(1 - \frac{n}{N}\right) \frac{S_d^2}{n}.
\]  (4.17)

Using the relation \(B_1 = RS_y/S_x\), it may be shown that

\[
\left(1 - \frac{n}{N}\right) S_d^2 = \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{i=1}^{N} \frac{(y_i - \bar{y}_U - B_1[x_i - \bar{x}_U])^2}{N - 1}
= \left(1 - \frac{n}{N}\right) \frac{1}{n} S_y^2(1 - R^2).
\]  (4.18)

(See Exercise 28.) Thus, the approximate MSE is small when

- \(n\) is large
- \(n/N\) is large
- \(S_y\) is small
- the correlation \(R\) is close to \(-1\) or \(+1\).

The standard error may be calculated by substituting estimates for the population quantities in (4.17) or (4.18). We can estimate \(S^2_d\) in (4.17) by using the residuals \(e_i = y_i - (\hat{B}_0 + \hat{B}_1x_i)\); then \(S^2_d = \sum_{i \in S} e_i^2 / (n - 1)\) estimates \(S^2_d\) and

\[
\text{SE}(\hat{\bar{y}}_{\text{reg}}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2_d}{n}}.
\]  (4.19)

In small samples, we may alternatively calculate \(s^2_e\) using the MSE from a regression analysis: \(s^2_e = \sum_{i \in S} e_i^2 / (n - 2)\). This adjusts the estimator for the degrees of freedom in the regression. To estimate the variance using the formulation in (4.18), substitute the sample variance \(s_y^2\) and the sample correlation \(r\) for the population quantities \(S_y^2\) and \(R\), obtaining

\[
\text{SE}(\hat{\bar{y}}_{\text{reg}}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{1}{n} s_y^2(1 - r^2)}.
\]  (4.20)

**Example 4.9**

To estimate the number of dead trees in an area, we divide the area into 100 square plots and count the number of dead trees on a photograph of each plot. Photo counts can be made quickly, but sometimes a tree is misclassified or not detected. So we select an SRS of 25 of the plots for field counts of dead trees. We know that the
The plot of photo and field tree-count data, along with the regression line. Note that \( \hat{\bar{y}}_{\text{reg}} \) is the predicted value from the regression equation when \( x = \bar{x}_U \). The point \((\bar{x}, \bar{y})\) is marked by “+” on the graph.

Population mean number of dead trees per plot from the photo count is 11.3. The data—plotted in Figure 4.5—are given below.

<table>
<thead>
<tr>
<th>Photo (x)</th>
<th>10</th>
<th>12</th>
<th>7</th>
<th>13</th>
<th>13</th>
<th>6</th>
<th>17</th>
<th>16</th>
<th>15</th>
<th>10</th>
<th>14</th>
<th>12</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field (y)</td>
<td>15</td>
<td>14</td>
<td>9</td>
<td>14</td>
<td>8</td>
<td>5</td>
<td>18</td>
<td>15</td>
<td>13</td>
<td>15</td>
<td>11</td>
<td>15</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Photo (x)</th>
<th>5</th>
<th>12</th>
<th>10</th>
<th>10</th>
<th>9</th>
<th>6</th>
<th>11</th>
<th>7</th>
<th>9</th>
<th>11</th>
<th>10</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field (y)</td>
<td>8</td>
<td>13</td>
<td>9</td>
<td>11</td>
<td>12</td>
<td>9</td>
<td>12</td>
<td>13</td>
<td>11</td>
<td>10</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

For these data, \( \bar{x} = 10.6, \bar{y} = 11.56, s_y^2 = 9.09 \), and the sample correlation between \( x \) and \( y \) is \( r = 0.62420 \). Fitting a straight line regression model gives

\[
\hat{y} = 5.059292 + 0.613274x
\]

with \( \hat{B}_0 = 5.059292 \) and \( \hat{B}_1 = 0.613274 \). In this example, \( x \) and \( y \) are positively correlated so that \( \bar{x} \) and \( \bar{y} \) are also positively correlated. Since \( \bar{x} < \bar{x}_U \), we expect that the sample mean \( \bar{y} \) is also too small; the regression estimate adds the quantity \( B_1(\bar{x}_U - \bar{x}) = 0.613(11.3 - 10.6) = 0.43 \) to \( \bar{y} \) to compensate.

Using (4.15), the regression estimate of the mean is

\[
\hat{\bar{y}}_{\text{reg}} = 5.059292 + 0.613274(11.3) = 11.99.
\]
From (4.20), the standard error is

\[ SE(\hat{\bar{y}}_{reg}) = \sqrt{\left(1 - \frac{25}{100}\right)(9.09)(1 - 0.62420^2)} = 0.408. \]

The standard error of \( \hat{\bar{y}}_{reg} \) is less than that for \( \bar{y} \):

\[ SE[\bar{y}] = \sqrt{\left(1 - \frac{25}{100}\right)s_y^2} = 0.522. \]

We expect regression estimation to increase the precision in this example because the variables photo and field are positively correlated. To estimate the total number of dead trees, use

\[ \hat{y}_{reg} = (100)(11.99) = 1199; \]

\[ SE(\hat{y}_{reg}) = (100)(0.408) = 40.8. \]

In SAS software, PROC SURVEYREG calculates regression estimates. Code for this example is given on the website; partial output is given below.

| Parameter       | Estimate   | Standard Error | t Value | Pr > |t|  | 95% Confidence Interval |
|-----------------|------------|----------------|---------|------|-----|-------------------------|
| Total field trees | 1198.92920 | 42.7013825     | 28.08   | <.0001 |   | 1110.79798 1287.06053    |
| Mean field trees | 11.98929   | 0.42700138     | 28.08   | <.0001 |   | 11.10791 12.87061       |

The standard errors given by SAS software are slightly larger than those obtained by hand calculation because SAS uses a slightly different estimator (see Section 11.7). In practice, we recommend using survey regression software for regression estimation to avoid roundoff errors.

### 4.3.2 Difference Estimation

**Difference estimation** is a special case of regression estimation, used when the investigator “knows” that the slope \( B_1 \) is 1. Difference estimation is often recommended in accounting when an SRS is taken. A list of accounts receivable consists of the book value for each account—the company’s listing of how much is owed on each account. In the simplest sampling scheme, the auditor scrutinizes a random sample of the accounts to determine the audited value—the actual amount owed—in order to estimate the error in the total accounts receivable. The quantities considered are:

\[ y_i = \text{audited value for company } i \]
\[ x_i = \text{book value for company } i. \]

Then, \( \bar{y} - \bar{x} \) is the mean difference for the audited accounts.
5. Cluster Sampling

5.1 Introduction and notations

- In general, we want the target and study populations to be the same. When they are not the same, the researcher must be careful to ensure that conclusions based on the sample results can be applied to the target population.

- Because of restrictions such as cost or scheduling conflicts, it is often impossible to collect a simple random sample or a stratified simple random sample. In many cases, however, it may be possible to define a sampling frame with sampling units that are not the units in the target population or the study population yet still obtain statistically valid estimates.
• Cluster sampling, and, specifically, systematic sampling are examples when a difference between the target population and the sampling frame occurs. Despite the difference, if executed properly, conclusions based on the sample results from these sampling designs can be applied to the target population.

• Situation: A population contains $M_0$ population units. The set of $M_0$ units is partitioned into $N$ disjoint groups of population units called primary sampling units (PSUs). The population units contained in the primary sampling units are called secondary sampling units (SSUs).

• The primary sampling units may be of different sizes. That is, the numbers of secondary sampling units in the primary sampling units are not all the same.
• The primary sampling units may be of different sizes. That is, the numbers of secondary sampling units in the primary sampling units are not all the same.

• The sampling units are clusters of population units. In this case, the sampling frame does not correspond with the units of the target population or the study population.

• Often, whenever any secondary sampling unit of a primary sampling unit is included in the sample, all the other secondary sampling units in that primary sampling unit will also be included in the sample.

• That is, the primary sampling units (PSUs) in the sampling frame for cluster sampling are strata. Typically, the number of strata is large, while each stratum contains only a small number of secondary sampling units (SSUs).
• Note: the population has $M_0$ individual units but the sampling frame has only $N$ primary sampling units corresponding the number of clusters (or strata) formed.

• The responses from the secondary sampling population units are not analyzed individually, but are combined with all other secondary sampling units that are in the same cluster. Therefore, there are $N$ possible $y$ values (not $M_0$).

• The researcher hopes that reducing the $M_0$ population units to a sampling frame containing only $N$ sampling units is offset by the practical conveniences (such as reduced cost) that this type of sampling frame can offer.
### F I G U R E  5.1

Similarities and differences between stratified sampling and one-stage cluster sampling

<table>
<thead>
<tr>
<th>Stratified Sampling</th>
<th>Cluster Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each element of the population is in exactly one stratum.</td>
<td>Each element of the population is in exactly one cluster.</td>
</tr>
<tr>
<td>Population of $H$ strata; stratum $h$ has $n_h$ elements:</td>
<td>One-stage cluster sampling; population of $N$ clusters:</td>
</tr>
<tr>
<td>![Stratified Sampling Diagram]</td>
<td>![Cluster Sampling Diagram]</td>
</tr>
<tr>
<td>Take an SRS from every stratum:</td>
<td>Take an SRS of clusters; observe all elements within the clusters in the sample:</td>
</tr>
<tr>
<td>![Stratified Sampling Diagram]</td>
<td>![Cluster Sampling Diagram]</td>
</tr>
<tr>
<td>Variance of the estimate of $\bar{Y}_U$ depends on the variability of values within strata.</td>
<td>The cluster is the sampling unit; the more clusters we sample, the smaller the variance. The variance of the estimate of $\bar{Y}_U$ depends primarily on the variability between cluster means.</td>
</tr>
<tr>
<td>For greatest precision, individual elements within each stratum should have similar values, but stratum means should differ from each other as much as possible.</td>
<td>For greatest precision, individual elements within each cluster should be heterogeneous, and cluster means should be similar to one another.</td>
</tr>
</tbody>
</table>

we partially repeat the same information instead of obtaining new information, and that gives us less precision for estimates of population quantities. Cluster sampling is used in practice because it is usually much cheaper and more convenient to sample in clusters than randomly in the population. Most large household surveys carried out by the U.S. government, or by commercial or academic institutions, use cluster sampling because of the cost savings.
5.1 Notation for Cluster Sampling

**psu Level—Population Quantities**

\[ N = \text{number of psus in the population} \]
\[ M_i = \text{number of ssus in psu } i \]
\[ M_0 = \sum_{i=1}^{N} M_i = \text{total number of ssus in the population} \]
\[ t_i = \sum_{j=1}^{M_i} y_{ij} = \text{total in psu } i \]
\[ t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \text{population total} \]
\[ S_i^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left( \frac{t_i - \frac{t}{N}}{M_0} \right)^2 = \text{population variance of the psu totals} \]

**ssu Level—Population Quantities**

\[ \bar{y}_U = \frac{1}{M_0} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0} = \text{population mean} \]
\[ \bar{y}_{iU} = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} = \text{population mean in psu } i \]
\[ S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_U)^2}{M_0 - 1} = \text{population variance (per ssu)} \]
\[ S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{iU})^2}{M_i - 1} = \text{population variance within psu } i \]

**Sample Quantities**

\[ n = \text{number of psus in the sample} \]
\[ m_i = \text{number of ssus in the sample from psu } i \]
\[ \bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m_i} = \text{sample mean (per ssu) for psu } i \]
\[ \hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij} = \text{estimated total for psu } i \]
\[ \hat{t}_{\text{unb}} = \sum_{i \in S} \frac{N}{n} \hat{t}_i = \text{unbiased estimator of population total} \]
\[ S_i^2 = \frac{1}{n-1} \sum_{i \in S} \left( \frac{\hat{t}_i - \hat{t}_{\text{unb}}}{N} \right)^2 \]
\[ s_i^2 = \sum_{j \in S_i} \frac{(y_{ij} - \bar{y}_i)^2}{m_i - 1} \quad \text{sample variance within psu } i \]
\[ w_{ij} = \text{sampling weight for ssu } j \text{ in psu } i \]

5.2 One-Stage Cluster Sampling

In one-stage cluster sampling, either all or none of the elements that compose a cluster (= psu) are in the sample. One-stage cluster sampling is used in many surveys in which the cost of sampling ssus is negligible compared with the cost of sampling psus. For education surveys, a natural psu is the classroom; all students in a selected classroom are often included as the ssus since little extra cost is added by handing out a questionnaire to all students in the classroom rather than some.

In the population of \(N\) psus, the \(i\)th psu contains \(M_i\) ssus (elements). In the simplest design, we take an SRS of \(n\) psus from the population and measure our variable of interest on every element in the sampled psus. Thus, for one-stage cluster sampling, \(M_i = m_i\).

5.2.1 Clusters of Equal Sizes: Estimation

Let’s consider the simplest case in which each psu has the same number of elements, with \(M_i = m_i = M\). Most naturally occurring clusters of people do not fit into this framework, but it can occur in agricultural and industrial sampling. Estimating population means or totals is simple: We treat the psu means or totals as the observations and simply ignore the individual elements.

Thus, we have an SRS of \(n\) data points \(\{t_i, i \in S\}\); \(t_i\) is the total for all the elements in psu \(i\). Then \(\bar{t}_S = \frac{\sum_{i \in S} t_i}{n}\) estimates the average of the cluster totals.

In a household survey to estimate income in two-person households, the individual observations \(y_{ij}\) are the incomes of individual persons within the household, \(t_i\) is the total income for household \(i\) (\(t_i\) is known for sampled households because both persons are interviewed), \(\bar{t}_U\) is the average income per household, and \(\bar{y}_U\) is the average income per person. To estimate the total income \(t\), we can use the estimator

\[ \hat{t} = \frac{N}{n} \sum_{i \in S} t_i. \] \hspace{1cm} (5.1)

The results in sections 2.3 and 2.8 apply to \(\hat{t}\) because we have an SRS of \(n\) units from a population of \(N\) units. As a result, \(\hat{t}\) is an unbiased estimator of \(t\), with variance given by

\[ V(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n} \] \hspace{1cm} (5.2)

and with standard error

\[ \text{SE}(\hat{t}) = N \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}}, \] \hspace{1cm} (5.3)
where $S_i^2$ and $s_i^2$ are the population and sample variance, respectively, of the psu totals:

$$S_i^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left( t_i - \frac{t}{N} \right)^2$$

and

$$s_i^2 = \frac{1}{n-1} \sum_{i \in S} \left( t_i - \frac{\hat{t}}{N} \right)^2.$$

To estimate $\hat{\bar{y}}_U$, divide the estimated total by the number of persons, obtaining

$$\hat{\bar{y}} = \frac{\hat{t}}{NM}, \quad (5.4)$$

with

$$V(\hat{\bar{y}}) = \left( 1 - \frac{n}{N} \right) \frac{S_i^2}{nM^2} \quad (5.5)$$

and

$$\text{SE}(\hat{\bar{y}}) = \frac{1}{M} \sqrt{\left( 1 - \frac{n}{N} \right) \frac{s_i^2}{n}}. \quad (5.6)$$

No new ideas are introduced to carry out one-stage cluster sampling; we simply use the results for simple random sampling with the psu totals as the observations.

**Example 5.2**

A student wants to estimate the average grade point average (GPA) in his dormitory. Instead of obtaining a listing of all students in the dorm and conducting an SRS, he notices that the dorm consists of 100 suites, each with four students; he chooses 5 of those suites at random, and asks every person in the 5 suites what her or his GPA is. The results are as follows:

<table>
<thead>
<tr>
<th>Person Number</th>
<th>Suite (psu)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>3.08</td>
<td>2.36</td>
<td>2.00</td>
<td>3.00</td>
<td>2.68</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2.60</td>
<td>3.04</td>
<td>2.56</td>
<td>2.88</td>
<td>1.92</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3.44</td>
<td>3.28</td>
<td>2.52</td>
<td>3.44</td>
<td>3.28</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>3.04</td>
<td>2.68</td>
<td>1.88</td>
<td>3.64</td>
<td>3.20</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>12.16</td>
<td>11.36</td>
<td>8.96</td>
<td>12.96</td>
<td>11.08</td>
</tr>
</tbody>
</table>

The psus are the suites, so $N = 100$, $n = 5$, and $M = 4$. The estimate of the population total (the estimated sum of all the GPAs for everyone in the dorm—a meaningless quantity for this example but useful for demonstrating the procedure) is

$$\hat{t} = \frac{100}{5} (12.16 + 11.36 + 8.96 + 12.96 + 11.08) = 1130.4.$$  

The average of the suite totals is estimated by $\bar{t} = 1130.4/100 = 11.304$, and

$$s_i^2 = \frac{1}{5-1} \left[ (12.16 - 11.304)^2 + \cdots + (11.08 - 11.304)^2 \right] = 2.256.$$
### 5.2 One-Stage Cluster Sampling

#### Data Summary

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Clusters</td>
<td>5</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>20</td>
</tr>
<tr>
<td>Sum of Weights</td>
<td>400</td>
</tr>
</tbody>
</table>

#### Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>gpa</td>
<td>20</td>
<td>2.826000</td>
<td>0.163665</td>
<td>2.37159339</td>
</tr>
</tbody>
</table>

If we had taken an SRS of \( nM \) elements, each element in the sample would have been assigned weight \( (NM)/(nM) = N/n \)—the same weights we obtain for cluster sampling. The precision obtained for the two types of sampling, however, can differ greatly; the difference in precision is explored in the next section.

### 5.2.2 Clusters of Equal Sizes: Theory

In this section we compare cluster sampling with simple random sampling: Cluster sampling almost always provides less precision for the estimators than one would obtain by taking an SRS with the same number of elements.

As in stratified sampling, let’s look at the ANOVA table (Table 5.1) for the whole population. In stratified sampling, the variance of the estimator of \( t \) depended on the variability within the strata; Equation (3.3) and Table 3.3 imply that the variance in stratified sampling is small if SSW is small relative to SSTO, or equivalently, if the within mean square (MSW) is small relative to \( S^2 \). In stratified sampling, you have some information about every stratum, so you need not worry about variability due to unsampled strata. If MSB/MSW is large—that is, the variability among the strata means is large when compared with the variability within strata—then stratified sampling increases precision.

The opposite situation occurs in cluster sampling. In one-stage cluster sampling when each psu has \( M \) ssus, the variability of the unbiased estimator of \( t \) depends entirely on the between-psu part of the variability, because

\[
\hat{s}_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(t_i - \bar{t}_U)^2}{M^2} = \frac{1}{N-1} \sum_{i=1}^{N} \frac{M^2(\bar{y}_iU - \bar{y}_U)^2}{N-1} = M(\text{MSB}).
\]

Thus, for cluster sampling,

\[
V(\hat{\mu}_{\text{cluster}}) = N^2 \left( 1 - \frac{n}{N} \right) \frac{M(\text{MSB})}{n}.
\]  

(5.7)

If MSB/MSW is large in cluster sampling, then cluster sampling decreases precision. In that situation, MSB is relatively large because it measures the cluster-to-cluster variability: Elements in different clusters often vary more than elements in the same cluster because different clusters have different means. If we took a cluster sample of classes and sampled all students within the selected classes, we would likely find
Table 5.1
Population ANOVA Table—Cluster Sampling

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between psus</td>
<td>(N - 1)</td>
<td>(\sum_{i=1}^{N} \sum_{j=1}^{M} (\bar{y}<em>{ijU} - \bar{y}</em>{U})^2)</td>
<td>MSB</td>
</tr>
<tr>
<td>Within psus</td>
<td>(N(M - 1))</td>
<td>(\sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_{ijU})^2)</td>
<td>MSW</td>
</tr>
<tr>
<td>Total, about (\bar{y}_{U})</td>
<td>(NM - 1)</td>
<td>(\sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_{U})^2)</td>
<td>(S^2)</td>
</tr>
</tbody>
</table>

that average reading scores varied from class to class. An excellent reading teacher might raise the reading scores for the entire class; a class of students from an area with much poverty might tend to be undernourished and not score as highly at reading. Unmeasured factors, such as teaching skill or poverty, can affect the overall mean for a cluster, and thus cause MSB to be large.

Within a class, too, students’ reading scores vary. The MSW is the pooled value of the within-cluster variances: the variance from element to element, present for all elements of the population. If the clusters are relatively homogeneous—if, for example, students in the same class have similar scores—the MSW will be small.

Now let’s compare cluster sampling to simple random sampling. If, instead of taking a cluster sample of \(M\) elements in each of \(n\) clusters, we had taken an SRS with \(nM\) observations, the variance of the estimated total would have been

\[
V(\hat{t}_{SRS}) = (NM)^2 \left(1 - \frac{nM}{NM}\right) \frac{S^2}{nM} = N^2 \left(1 - \frac{n}{N}\right) \frac{M S^2}{n}.
\]

Comparing this with (5.7), we see that if MSB > \(S^2\), then cluster sampling is less efficient than simple random sampling.

The intraclass (sometimes called intracluster) correlation coefficient (ICC) tells us how similar elements in the same cluster are. It provides a measure of homogeneity within the clusters. ICC is defined to be the Pearson correlation coefficient for the \(NM(M - 1)\) pairs \((y_{ij}, y_{ik})\) for \(i\) between 1 and \(N\) and \(j \neq k\) (see Exercise 22) and can be written in terms of the population ANOVA table quantities as

\[
ICC = 1 - \frac{M}{M - 1} \frac{SSW}{SSTO}.
\]

Because \(0 \leq SSW/SSTO \leq 1\), it follows from (5.8) that

\[
- \frac{1}{M - 1} \leq ICC \leq 1.
\]
5.2 One-Stage Cluster Sampling

If the clusters are perfectly homogeneous and hence SSW = 0, then ICC = 1. Equation (5.8) also implies that

\[ MSB = \frac{NM - 1}{M(N - 1)} S^2 [1 + (M - 1)ICC]. \]  

(5.9)

How much precision do we lose by taking a cluster sample? From (5.7) and (5.9),

\[ \frac{V(\hat{t}_{\text{cluster}})}{V(\hat{t}_{\text{SRS}})} = \frac{MSB}{S^2} = \frac{NM - 1}{M(N - 1)} [1 + (M - 1)ICC]. \]  

(5.10)

If \( N \), the number of psus in the population, is large so that \( NM - 1 \approx M(N - 1) \), then the ratio of the variances in (5.10) is approximately \( 1 + (M - 1)ICC \). So \( 1 + (M - 1)ICC \) ssus, taken in a one-stage cluster sample, give us approximately the same amount of information as one ssu from an SRS. If ICC = 1/2 and M=5, then \( 1 + (M - 1)ICC = 3 \), and we would need to measure 300 elements using a cluster sample to obtain the same precision as an SRS of 100 elements. We hope, though, that because it is often much cheaper and easier to collect data in a cluster sample, that we will have more precision per dollar spent in cluster sampling.

The ICC provides a measure of homogeneity for the clusters. The ICC is positive if elements within a psu tend to be similar; then, SSW will be small relative to SSTO, and the ICC relatively large. When the ICC is positive, cluster sampling is less efficient than simple random sampling of elements.

If the clusters occur naturally in the population, the ICC is usually positive. Elements within the same cluster tend to be more similar than elements selected at random from the population. This may occur because the elements in a cluster share a similar environment—we would expect wells in the same geographic cluster to have similar levels of pesticides, or we would expect one area of a city to have a different incidence of measles than another area of a city. In human populations, personal choice as well as interactions among household members or neighbors may cause the ICC to be positive—wealthy households tend to live in similar neighborhoods, and persons in the same neighborhood may share similar opinions.

The ICC is negative if elements within a cluster are dispersed more than a randomly chosen group would be. This forces the cluster means to be very nearly equal—because SSTO = SSW + SSB, if SSTO is held fixed and SSW is large, then SSB must be small. If ICC < 0, cluster sampling is more efficient than simple random sampling of elements. The ICC is rarely negative in naturally occurring clusters, but negative values can occur in some systematic samples or artificial clusters, as discussed in Section 5.5.

The ICC is only defined for clusters of equal sizes. An alternative measure of homogeneity in general populations is the adjusted \( R^2 \), called \( R_a^2 \) and defined as

\[ R_a^2 = 1 - \frac{MSW}{S^2}. \]  

(5.11)

If all psus are of the same size, then the increase in variance due to cluster sampling is

\[ \frac{V(\hat{t}_{\text{cluster}})}{V(\hat{t}_{\text{SRS}})} = \frac{MSB}{S^2} = 1 + \frac{N(M - 1)}{N - 1} R_a^2, \]
by comparing with (5.10), you can see that for many populations, \( R^2_a \) is close to the ICC. The quantity \( R^2_a \) is a reasonable measure of homogeneity because of its interpretation in linear regression: It is the relative amount of variability in the population explained by the psu means, adjusted for the number of degrees of freedom. If the psus are homogeneous, then the psu means are highly variable relative to the variation within psus, and \( R^2_a \) will be high.

**Example 5.3** Consider two artificial populations, each having three psus with three elements per psu.

<table>
<thead>
<tr>
<th></th>
<th>Population A</th>
<th>Population B</th>
</tr>
</thead>
<tbody>
<tr>
<td>psu 1</td>
<td>10 20 30</td>
<td>9 10 11</td>
</tr>
<tr>
<td>psu 2</td>
<td>11 20 32</td>
<td>17 20 20</td>
</tr>
<tr>
<td>psu 3</td>
<td>9 17 31</td>
<td>31 32 30</td>
</tr>
</tbody>
</table>

The elements are the same in the two populations, so the populations share the values \( \bar{y}_U = 20 \) and \( S^2 = 84.5 \). In population A, the psu means are similar and most of the variability occurs within psus; in population B, most of the variability occurs between psus.

<table>
<thead>
<tr>
<th></th>
<th>Population A</th>
<th>Population B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \bar{y}_i )</td>
<td>( S_i^2 )</td>
</tr>
<tr>
<td>psu 1</td>
<td>20 100</td>
<td>10 1</td>
</tr>
<tr>
<td>psu 2</td>
<td>21 111</td>
<td>19 3</td>
</tr>
<tr>
<td>psu 3</td>
<td>19 124</td>
<td>31 1</td>
</tr>
</tbody>
</table>

ANOVA Table for Population A:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between psus</td>
<td>2</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Within psus</td>
<td>6</td>
<td>670</td>
<td>111.67</td>
</tr>
<tr>
<td>Total, about mean</td>
<td>8</td>
<td>676</td>
<td>84.5</td>
</tr>
</tbody>
</table>

For population A:

\[
R^2_a = 1 - \frac{111.67}{84.5} = -0.3215
\]

\[
\text{ICC} = 1 - \left( \frac{3}{2} \right) \frac{670}{676} = -0.4867
\]

ANOVA Table for Population B:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between psus</td>
<td>2</td>
<td>666</td>
<td>333</td>
</tr>
<tr>
<td>Within psus</td>
<td>6</td>
<td>10</td>
<td>1.67</td>
</tr>
<tr>
<td>Total, about mean</td>
<td>8</td>
<td>676</td>
<td>84.5</td>
</tr>
</tbody>
</table>

For population B:

\[
R^2_a = 1 - \frac{1.67}{84.5} = 0.9803
\]

\[
\text{ICC} = 1 - \left( \frac{3}{2} \right) \frac{10}{676} = 0.9778
\]

Population A has much variation among elements within the psus, but little variation among the psu means. This is reflected in the large negative values of the ICC and \( R^2_a \): Elements in the same cluster are actually less similar than randomly selected elements from the whole population. For this situation, cluster sampling is more efficient than simple random sampling.
In a one-stage cluster sample of \( n \) of the \( N \) psus, we know how to estimate population totals and means in two ways: using unbiased estimation and using ratio estimation.

**Unbiased Estimation.** An *unbiased* estimator of \( t \) is calculated exactly as in (5.1):

\[
\hat{t}_{\text{unb}} = \frac{N}{n} \sum_{i \in S} t_i,
\]

and, by (5.3),

\[
\text{SE}(\hat{t}_{\text{unb}}) = N \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2_t}{n}}.
\]

The difference between unequal- and equal-sized clusters is that the variation among the individual cluster totals \( t_i \) is likely to be large when the clusters have different sizes. The investigators conducting the Enumerative Check Census of 1937 were interested in the total number of unemployed persons, and \( t_i \) would be the number of unemployed persons in postal route \( i \). One would expect to find more persons, and hence more unemployed persons, on a postal route with a large number of households than on a postal route with a small number of households. So we would expect that \( t_i \) would be large when the psu size \( M_i \) is large, and small when \( M_i \) is small. Often, then, \( s^2_t \) is larger in a cluster sample when the psus have unequal sizes than when the psus all have the same number of ssus.

The probability that a psu is in the sample is \( n/N \), as an SRS of \( n \) of the \( N \) psus is taken. Since one-stage cluster sampling is used, an ssu is included in the sample whenever its psu is included in the sample. Thus, as in Section 5.2.1,

\[
w_{ij} = \frac{1}{P\{\text{ssu } j \text{ of psu } i \text{ is in sample}\}} = \frac{N}{n}.
\]

One-stage cluster sampling produces a self-weighting sample when the psus are selected with equal probabilities. Using the weights, (5.12) may be written as

\[
\hat{t}_{\text{unb}} = \sum_{i \in S} \sum_{j \in S_i} w_{ij}y_{ij}.
\]

We can use (5.12) and (5.13) to derive an unbiased estimator for \( \bar{y}_U \) and to find its standard error. Define

\[
M_0 = \sum_{i=1}^{N} M_i
\]

as the total number of ssus in the population; then \( \hat{\bar{y}}_{\text{unb}} = \hat{t}_{\text{unb}}/M_0 \) and \( \text{SE}(\hat{\bar{y}}_{\text{unb}}) = \text{SE}(\hat{t}_{\text{unb}})/M_0 \). The unbiased estimator of the mean \( \bar{y}_{\text{unb}} \) can be inefficient when the values of \( M_i \) are unequal since it, like \( \hat{t}_{\text{unb}} \), depends on the variability of the cluster totals \( t_i \). It also requires that \( M_0 \) be known; however, we often know \( M_i \) only for the sampled clusters. In the Enumerative Check Census, for example, the number of households on a postal route would only be ascertained for the postal routes actually chosen to be in the sample. We now examine another estimator for \( \hat{\bar{y}}_U \) that is usually more efficient when the population psu sizes are unequal.
**Chapter 5: Cluster Sampling with Equal Probabilities**

**Ratio Estimation.** We usually expect \( t_i \) to be positively correlated with \( M_i \). If psus are counties, we would expect the total number of households living in poverty in county \( i \) (\( t_i \)) to be roughly proportional to the total number of households in county \( i \) (\( M_i \)). The population mean \( \hat{y}_U \) is a ratio

\[
\hat{y}_U = \frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{N} M_i} = \frac{t}{M_0},
\]

where \( t_i \) and \( M_i \) are usually positively correlated. Thus, \( \hat{y}_U = B \) as in Section 4.1 (substituting \( t_i \) for \( y_i \) and using \( M_i \) as the auxiliary variable \( x_i \)). Define

\[
\hat{y}_r = \frac{\hat{y}_{\text{unb}}}{M_0} = \frac{\sum_{i \in S} t_i}{\sum_{i \in S} M_i},
\]

where

\[
\hat{y}_{\text{unb}} = \frac{\sum_{i \in S, j \in S_i} w_{ij} y_{ij}}{\sum_{i \in S} \sum_{j \in S_i} w_{ij}}.
\]

(5.15)

Note that \( \hat{y}_r \) from (5.15) may also be calculated using the weights \( w_{ij} \), as

\[
\hat{y}_r = \frac{\hat{y}_{\text{unb}}}{M_0} = \frac{\sum_{i \in S, j \in S_i} w_{ij} y_{ij}}{\sum_{i \in S} \sum_{j \in S_i} w_{ij}}.
\]

(5.16)

Since an SRS of clusters is selected, all the weights are the same with \( w_{ij} = N/n \).

The estimator \( \hat{y}_r \) in (5.15) is the quantity \( \hat{B} \) in (4.2): the denominator is a random quantity that depends on which particular psus are included in the sample. If the \( M_i \)'s are unequal and a different cluster sample of size \( n \) is taken, the denominator will likely be different. From (4.10),

\[
\text{SE}(\hat{y}_r) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{1}{nM^2} \frac{\sum_{i \in S} (t_i - \hat{y}_r M_i)^2}{n - 1}}
\]

\[
\text{SE}(\hat{y}_r) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{1}{nM^2} \frac{\sum_{i \in S} M_i^2 (\hat{y}_i - \hat{y}_r)^2}{n - 1}}.
\]

(5.17)

The variance of the ratio estimator depends on the variability of the means per element in the clusters, and can be much smaller than that of the unbiased estimator \( \hat{y}_{\text{unb}} \).

If the total number of elements in the population, \( M_0 = \sum_{i=1}^{N} M_i \), is known, we can also use ratio estimation to estimate the population total: the ratio estimator is \( \hat{t}_r = M_0 \hat{y}_r \) with \( \text{SE}(\hat{t}_r) = M_0 \text{SE}(\hat{y}_r) \). Note, though, that \( \hat{t}_r \) requires that we know the total number of elements in the population, \( M_0 \); the unbiased estimator in (5.12) makes no such requirement.
**Example 5.6** One-stage cluster samples are often used in educational studies, since students are naturally clustered into classrooms or schools. Consider a population of 187 high school algebra classes in a city. An investigator takes an SRS of 12 of those classes and gives each student in the sampled classes a test about function knowledge. The (hypothetical) data are given in the file algebra.dat, with the following summary statistics.

<table>
<thead>
<tr>
<th>Class Number</th>
<th>( M_i )</th>
<th>( \bar{y}_i )</th>
<th>( t_i )</th>
<th>( M^2_i(\bar{y}_i - \hat{\bar{y}}_r)^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>20</td>
<td>61.5</td>
<td>1,230</td>
<td>456.7298</td>
</tr>
<tr>
<td>37</td>
<td>26</td>
<td>64.2</td>
<td>1,670</td>
<td>1,867.7428</td>
</tr>
<tr>
<td>38</td>
<td>24</td>
<td>58.4</td>
<td>1,402</td>
<td>9,929.2225</td>
</tr>
<tr>
<td>39</td>
<td>34</td>
<td>58.0</td>
<td>1,972</td>
<td>24,127.7518</td>
</tr>
<tr>
<td>41</td>
<td>26</td>
<td>58.0</td>
<td>1,508</td>
<td>14,109.3082</td>
</tr>
<tr>
<td>44</td>
<td>28</td>
<td>64.9</td>
<td>1,816</td>
<td>4,106.2808</td>
</tr>
<tr>
<td>46</td>
<td>19</td>
<td>55.2</td>
<td>1,048</td>
<td>19,825.3937</td>
</tr>
<tr>
<td>51</td>
<td>32</td>
<td>72.1</td>
<td>2,308</td>
<td>93,517.3218</td>
</tr>
<tr>
<td>58</td>
<td>17</td>
<td>58.2</td>
<td>989</td>
<td>5,574.9446</td>
</tr>
<tr>
<td>62</td>
<td>21</td>
<td>66.6</td>
<td>1,398</td>
<td>7,066.1174</td>
</tr>
<tr>
<td>106</td>
<td>26</td>
<td>62.3</td>
<td>1,621</td>
<td>33,4386</td>
</tr>
<tr>
<td>108</td>
<td>26</td>
<td>67.2</td>
<td>1,746</td>
<td>142,127.867</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>299</td>
<td>18,708</td>
<td>194,827.0387</td>
<td></td>
</tr>
</tbody>
</table>

We can use either (5.15) or (5.16) to estimate the mean score in the population: Using (5.15),

\[
\hat{\bar{y}}_r = \frac{\sum_{i \in S} M_i \bar{y}_i}{\sum_{i \in S} M_i} = \frac{18,708}{299} = 62.57.
\]

The standard error, from (5.17), is

\[
SE(\hat{\bar{y}}_r) = \sqrt{\left(1 - \frac{12}{187}\right) \frac{1}{(12)(24.92^2)}} \frac{194,827}{11} = 1.49.
\]

The weight for each observation is \( w_{ij} = 187/12 = 15.5833 \); we can alternatively calculate \( \hat{\bar{y}}_r \) using (5.16) as

\[
\hat{\bar{y}}_r = \frac{\sum_{i \in S} \sum_{j=1}^{M_i} w_{ij} y_{ij}}{\sum_{i \in S} \sum_{j=1}^{M_i} w_{ij}} = \frac{291,533}{4659.41667} = 62.57.
\]

SAS software uses (5.16) to estimate \( \bar{y}_U \). SAS code for calculating these estimates and producing the following output is given on the website.
### Data Summary

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Clusters</td>
<td>12</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>299</td>
</tr>
<tr>
<td>Sum of Weights</td>
<td>4659.41667</td>
</tr>
</tbody>
</table>

### Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>DF</th>
<th>Mean of Mean</th>
<th>Std Error of Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>score</td>
<td>299</td>
<td>11</td>
<td>62.568562</td>
<td>1.491578</td>
</tr>
</tbody>
</table>

The sum of the weights for the sample, 4659.41667, estimates the total number of students in the 187 high school algebra classes.

## 5.3 Two-Stage Cluster Sampling

In one-stage cluster sampling, we observe all the ssus within the selected psus. In many situations, though, the elements in a cluster may be so similar that measuring all subunits within a psu wastes resources; alternatively, it may be expensive to measure ssus relative to the cost of sampling psus. In these situations, it may be much cheaper to take a subsample within each psu selected. The stages within a two-stage cluster sample, when we sample the psus and subsample the ssus with equal probabilities, are:

1. Select an SRS $S$ of $n$ psus from the population of $N$ psus.
2. Select an SRS of ssus from each selected psu. The SRS of $m_i$ elements from the $i$th psu is denoted $S_i$.

The difference between one-stage and two-stage cluster sampling is illustrated in Figure 5.2. The extra stage complicates the notation and estimators, as we need to consider variability arising from both stages of data collection. The point estimators of $t$ and $\bar{y}_U$ are analogous to those in one-stage cluster sampling, but the variance formulas become messier.

In one-stage cluster sampling, we could estimate the population total by $\hat{t}_{\text{unb}} = (N/n) \sum_{i \in S} t_i$; the psu totals $t_i$ were known because we sampled every ssu in the selected psus. In two-stage cluster sampling, however, since we do not observe every ssu in the sampled psus, we need to estimate the individual psu totals by

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_j}{m_i} \bar{y}_{ij} = M_i \bar{y}_i$$
FIGURE 5.2
The difference between one-stage and two-stage cluster sampling.

One-Stage | Two-Stage
---|---
Population of $N$ psu's: & Population of $N$ psu's: 
Take an SRS of $n$ psu's: & Take an SRS of $n$ psu's: 
Sample all ssu's in sampled psu's: & Take an SRS of $m_i$ ssu's in sampled psu $i$: 

and an unbiased estimator of the population total is

$$
\hat{t}_{\text{unb}} = \frac{N}{n} \sum_{i \in S} \hat{t}_i = \frac{N}{n} \sum_{i \in S} M_i \bar{y}_i = \sum_{i \in S} \sum_{j \in S_i} \frac{N}{m_i} M_i \bar{y}_{ij}.
$$

(5.18)
For estimating means and totals in cluster samples, most survey statisticians use sampling weights. Equation (5.18) suggests that the sampling weight for ssu \( j \) of psu \( i \) is \( \frac{N M_i}{n m_i} \), and we can see that this is so by calculating the inclusion probability. For cluster sampling,

\[
P( \text{\( j \)th ssu in \( i \)th psu is selected}) = P(\text{\( i \)th psu selected}) \times P(\text{\( j \)th ssu selected \mid \text{\( i \)th psu selected}) = \frac{n m_i}{N M_i}.
\]

Recall from Section 2.4 that the weight of an element is the reciprocal of the probability of its selection. Thus,

\[
w_{ij} = \frac{NM_i}{nm_i} \tag{5.19}
\]

If psus are blocks, for example, and ssus are households, then household \( j \) in psu \( i \) represents \( (NM_i)/(nm_i) \) households in the population: itself, and \( (NM_i)/(nm_i) − 1 \) households that are not sampled. Then,

\[
\hat{t}_{\text{unb}} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}. \tag{5.20}
\]

In two-stage cluster sampling, a self-weighting design has each ssu representing the same number of ssus in the population. To take a cluster sample of persons in Illinois, we could take an SRS of counties in Illinois and then take an SRS of \( m_i \) of the \( M_i \) persons from county \( i \) in the sample. To have every person in the sample represent the same number of persons in the population, \( m_i \) needs to be proportional to \( M_i \), so that \( m_i/M_i \) is approximately constant. Thus, we would subsample more persons in the large counties than in the small counties to have a self-weighting sample.

The sampling weights provide a convenient way of calculating point estimates; they do not avoid associated shortcomings such as large variances. Also, the sampling weights give no information on how to find standard errors: We need to derive the formula for the variance using the sampling design.

In two-stage sampling, the \( \hat{t}_i \)'s are random variables. Consequently, the variance of \( \hat{t}_{\text{unb}} \) has two components: (1) the variability between psus and (2) the variability of ssus within psus. We do not have to worry about component (2) in one-stage cluster sampling.

The variance of \( \hat{t}_{\text{unb}} \) in (5.18) equals the variance of \( \hat{t}_{\text{unb}} \) from one-stage cluster sampling plus an extra term to account for the extra variance due to estimating the \( \hat{t}_i \)'s rather than measuring them directly. For two-stage cluster sampling,

\[
V(\hat{t}_{\text{unb}}) = N^2 \left( 1 - \frac{n}{N} \right) \frac{S_i^2}{n} + \frac{N}{n} \sum_{i=1}^{N} \left( 1 - \frac{m_i}{M_i} \right) M_i^2 \frac{S_i^2}{m_i}, \tag{5.21}
\]

where \( S_i^2 \) is the population variance of the cluster totals, and \( S_i^2 \) is the population variance among the elements within cluster \( i \). The first term in (5.21) is the variance from one-stage cluster sampling, and the second term is the additional variance due to subsampling within the psus. If \( m_i = M_i \) for each psu \( i \), as occurs in one-stage cluster sampling, then the second term in (5.21) is 0. To prove (5.21), we need to condition
5.3 Two-Stage Cluster Sampling

on the units included in the sample. This is more easily done in the general setting of unequal probability sampling; to avoid proving the same result twice, we shall prove the general result in Section 6.6.¹

To estimate $V(\hat{t}_{\text{unb}})$, let

$$s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left( \hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N} \right)^2$$

be the sample variance among the estimated psu totals and let

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j \in S_i} \left( y_{ij} - \bar{y}_i \right)^2$$

be the sample variance of the ssus sampled in psu $i$. As will be shown in Section 6.6, an unbiased estimator of the variance in (5.21) is given by

$$\hat{V}(\hat{t}_{\text{unb}}) = N^2 \left( 1 - \frac{n}{N} \right) S_t^2 \frac{1}{n} + \frac{N}{n} \sum_{i \in S} \left( 1 - \frac{m_i}{M_i} \right) M_i^2 S_i^2 \frac{1}{m_i}.$$  (5.24)

The standard error, $SE(\hat{t}_{\text{unb}})$, is of course the square root of (5.24).

**Remark.** In many situations when $N$ is large, the contribution of the second term in (5.24) to the variance estimator is negligible compared with that of the first term. We show in Section 6.6 that

$$E[s_t^2] = S_t^2 + \frac{1}{N} \sum_{i=1}^N \left( 1 - \frac{m_i}{M_i} \right) M_i^2 S_i^2 \frac{1}{m_i}.$$  

We expect the sample variance of the estimated psu totals $\hat{t}_i$ to be larger than the sample variance of the true psu totals $t_i$ because $\hat{t}_i$ will be different if we take a different subsample in psu $i$. Thus, if $N$ is large, the first term in (5.24) is approximately unbiased for the theoretical variance in (5.21). To simplify calculations, most software packages for analyzing survey data (including SAS software) estimate the variance using only the first term of (5.24), often omitting the finite population correction (fpc), $(1 - n/N)$. The estimator

$$\hat{V}_{WR}(\hat{t}_{\text{unb}}) = N^2 S_t^2 \frac{1}{n}$$  (5.25)

estimates the with-replacement variance for a cluster sample, as will be seen in Section 6.3. If the first-stage sampling fraction $n/N$ is small, there is little difference between the variance from a with-replacement sample and that from a without-replacement sample. Alternatively, a replication method of variance estimation from Chapter 9 can be used.

If we know the total number of elements in the population, $M_0$, we can estimate the population mean by $\hat{\bar{y}}_{\text{unb}} = \hat{t}_{\text{unb}} / M_0$ with standard error $SE(\hat{\bar{y}}_{\text{unb}}) = SE(\hat{t}_{\text{unb}}) / M_0$.

¹Working with the additional level of abstraction will allow us to see the structure of the variance more clearly, without floundering in the notation of the special case of equal probabilities discussed in this chapter. If you prefer to see the proof before you use the variance results, read Section 6.6 now.
As in one-stage cluster sampling with unequal cluster sizes, $s_t^2$ can be very large since it is affected both by variations in the unit sizes ($M_i$) and by variations in the $\bar{y}_i$. If the cluster sizes are disparate, this component is large, even if the cluster means are fairly constant.

**Ratio Estimation.** As in one-stage cluster sampling, we use a ratio estimator for the population mean. Again, the $y$’s of Chapter 4 are the psu totals (now estimated by $\hat{t}_i$) and the $x$’s are the psu sizes $M_i$. As in (5.15),

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \hat{t}_i}{\sum_{i \in S} M_i} = \frac{\sum_{i \in S} M_i \bar{y}_i}{\sum_{i \in S} M_i}. \tag{5.26}$$

Using the sampling weights in (5.19) with $w_{ij} = \frac{(NM_i)}{(nm_i)}$, we can rewrite $\hat{\bar{y}}_r$ as

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}}{\sum_{i \in S} \sum_{j \in S_i} w_{ij}}. \tag{5.27}$$

The weights are different, but the form of the estimator is the same as in (5.16). The variance estimator is again based on the approximation in (4.10):

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\hat{M}^2} \left( \frac{1}{n} - \frac{1}{N} \right) \frac{s_t^2}{n} + \frac{1}{nN\hat{M}^2} \sum_{i \in S} M_i^2 \left( 1 - \frac{m_i}{M_i} \right) \frac{s_i^2}{m_i}, \tag{5.28}$$

where $s_i^2$ is defined in (5.23),

$$s_r^2 = \frac{1}{n-1} \sum_{i \in S} (M_i \bar{y}_i - M_i \hat{\bar{y}}_r)^2, \tag{5.29}$$

and $\hat{M}$ is the average psu size. As with $\hat{t}_{unb}$, the second term in (5.28) is usually negligible compared with the first term, and most survey software packages calculate the variance using only the first term.

**Example 5.7**

The data in the file coots.dat come from Arnold’s (1991) work on egg size and volume of American Coot eggs in Minnedosa, Manitoba. In this data set, we look at volumes of a subsample of eggs in clutches (nests of eggs) with at least two eggs available for measurement.

The data are plotted in Figures 5.3 and 5.4. Data from a cluster sample can be plotted in many ways, and you often need to construct more than one type of plot to see features of the data. Because we have only two observations per clutch, we can plot the individual data points. If we had many observations per clutch, we could instead construct side-by-side boxplots, with one boxplot for each psu (we did a similar plot in Figure 3.1 for a stratified sample, constructing a boxplot for each stratum). We shall return to the issue of plotting data from complex surveys in Section 7.4.

Next, we use a spreadsheet (partly shown in Table 5.2; the full spreadsheet is on the website) to calculate summary statistics for each clutch. The summary statistics may then be used to estimate the average egg volume and its variance. The numbers
5.3 Two-Stage Cluster Sampling

**Figure 5.3**
Plot of egg volume data. Note the wide variation in the means from clutch to clutch. This indicates that eggs within the same clutch tend to be more similar than two randomly selected eggs from different clutches, and that clustering does not provide as much information per egg as would an SRS of eggs.

**Figure 5.4**
Another plot of egg volume data. Here, we ordered the clutches from smallest mean to largest mean, and drew the line connecting the two measurements of volume from the eggs in the clutch. Clutch number 88, represented by the long line in the middle of the graph, has an unusually large difference between the two eggs: One egg has volume 1.85, and the other has volume 2.84.
TABLE 5.2  
Part of Spreadsheet Used for Calculations in Example 5.7

<table>
<thead>
<tr>
<th>Clutch</th>
<th>$M_i$</th>
<th>$\hat{y}_i$</th>
<th>$s^2_i$</th>
<th>$\hat{t}_i$</th>
<th>$(1 - \frac{2}{M_i})M_i^2s^2_i/M_i$</th>
<th>$(\hat{t}_i - M_i\hat{\bar{y}}_r)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>3.86</td>
<td>0.0094</td>
<td>50.23594</td>
<td>0.671901</td>
<td>318.9232</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>4.19</td>
<td>0.0009</td>
<td>54.52438</td>
<td>0.065615</td>
<td>490.4832</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.92</td>
<td>0.0005</td>
<td>5.49750</td>
<td>0.005777</td>
<td>89.2263</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>3.00</td>
<td>0.0008</td>
<td>32.98168</td>
<td>0.039354</td>
<td>31.19576</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>182</td>
<td>13</td>
<td>4.22</td>
<td>0.0003</td>
<td>54.85854</td>
<td>0.002625</td>
<td>505.3962</td>
</tr>
<tr>
<td>183</td>
<td>13</td>
<td>4.41</td>
<td>0.0088</td>
<td>57.39262</td>
<td>0.630563</td>
<td>625.7549</td>
</tr>
<tr>
<td>184</td>
<td>12</td>
<td>3.48</td>
<td>0.00006</td>
<td>41.81168</td>
<td>0.000400</td>
<td>142.1994</td>
</tr>
<tr>
<td>sum</td>
<td>1757</td>
<td></td>
<td></td>
<td>4375.94652</td>
<td>42.174452</td>
<td>11,439.5794</td>
</tr>
<tr>
<td>$\hat{\bar{y}}_r = $</td>
<td></td>
<td></td>
<td></td>
<td>2.490579</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The values in the table have been rounded so that they fit on the page; in practice, of course, you should carry out all calculations to machine precision.

We use the ratio estimator to estimate the mean egg volume. From (5.26),

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \hat{t}_i}{\sum_{i \in S} M_i} = \frac{4375.947}{1757} = 2.49.$$ 

From the spreadsheet (Table 5.2),

$$s^2_r = \frac{1}{n - 1} \sum_{i \in S} (\hat{t}_i - M_i\hat{\bar{y}}_r)^2 = \frac{11,439.58}{183} = 62.51$$

and $\bar{M}_S = 1757/184 = 9.549$. Using (5.28), then,

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{9.549^2} \left( \left( 1 - \frac{184}{N} \right) \frac{62.511}{184} + \frac{1}{N} \frac{42.17}{184} \right).$$

Now $N$, the total number of clutches in the population, is unknown but presumed to be large (and known to be larger than 184). Thus, we may take the psu-level fpc to be 1, and note that the second term in the estimated variance will be very small relative to the first term. We then use

$$SE(\hat{\bar{y}}_r) = \frac{1}{9.549} \sqrt{\frac{62.511}{184}} = 0.061.$$

The estimated coefficient of variation for $\hat{\bar{y}}_r$ is

$$\frac{SE(\hat{\bar{y}}_r)}{\hat{\bar{y}}_r} = \frac{0.061}{2.49} = 0.0245.$$
7.4 Systematic Sampling

- **Systematic sampling** is a sampling plan in which the population units are collected systematically throughout the population. More specifically, a single primary sampling unit consists of secondary sampling units that are relatively spaced with each other in some systematic pattern throughout the population.

- Suppose the study area is partitioned into a $20 \times 20$ grid of 400 population units. A primary sampling unit in a systematic sample could consist of all population units that form a lattice which are 5 units apart horizontally and vertically. In Figure 9a, $N = 25$ and $M = 16$. In Figure 9b, each of the $N = 50$ primary sampling units contains $M = 8$ secondary sampling units.

- Initially, systematic sampling and cluster sampling appear to be opposites because systematic samples contain secondary sampling units that are spread throughout the population (good global coverage of the study area) while cluster samples are collected in groups of close proximity (good coverage locally within the study area).

- Systematic and cluster sampling are similar, however, because whenever a primary sampling unit is selected from the sampling frame, all secondary sampling units of that primary sampling unit will be included in the sample. Thus, random selection occurs at the primary sampling unit level and not the secondary sampling unit level.

- For estimation purposes, you could ignore the secondary sampling unit $y_{ij}$-values and only retain the primary sampling units $t_i$-values. This is what we did with one-stage cluster sampling.

- The **systematic and cluster sampling principle**: To obtain estimators of low variance, the population must be partitioned into primary sampling unit clusters in such a way that the clusters are similar to each other with respect to the $t_i$-values (small cluster-to-cluster variability).
• This is equivalent to saying that the within-cluster variability should be as large as possible to obtain the most precise estimators. Thus, the ideal primary sampling unit is representative of the full diversity of \( y_{ij} \)-values within the population.

• With natural populations of spatially distributed plants, animals, minerals, etc., these conditions are typically satisfied by systematic primary sampling units (and are not satisfied by primary sampling units with spatially clustered secondary sampling units).

7.4.1 Estimation of \( \bar{y}_U \) and \( t \)

• If a SRS is used to select the systematic primary sampling units, we can apply the estimation results for cluster sampling to define (i) estimators, (ii) the variance of each estimator, and (iii) the estimated variance of each estimator.

• The following formulas will be the same as those used for one-stage cluster sampling. The subscript \( \text{sys} \) denotes the fact that data were collected under systematic sampling.

\[
\hat{t}_{\text{sys}} = \frac{N}{n} \sum_{i=1}^{n} t_i = N \bar{y} \\
\hat{\bar{y}}_{U\text{sys}} = \frac{1}{nM} \sum_{i=1}^{n} t_i = \frac{\bar{y}}{M} = \frac{\hat{t}_{\text{sys}}}{M_0} 
\]

with variance

\[
V(\hat{t}_{\text{sys}}) = \frac{N(N-n)}{n} \frac{S_t^2}{n} \\
V(\hat{\bar{y}}_{U\text{sys}}) = \frac{N(N-n)}{M_0^2} \frac{S_t^2}{n} 
\]

where \( S_t^2 = \frac{\sum_{i=1}^{N} (t_i - \bar{t}_i)^2}{N-1} \).

• Recall that \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} t_i \) is the sample mean and that \( s_t^2 = \frac{\sum_{i=1}^{n} (t_i - \bar{y})^2}{n-1} \) is the sample variance of the primary sampling units.

• Because \( S_t^2 \) is unknown, we use \( s_u^2 \) to get unbiased estimators of the variances:

\[
\hat{V}(\hat{t}_{\text{sys}}) = \frac{N(N-n)}{n} \frac{s_t^2}{n} \\
\hat{V}(\hat{\bar{y}}_{U\text{sys}}) = \frac{N(N-n)}{M_0^2} \frac{s_t^2}{n} 
\]

7.4.2 Confidence Intervals for \( \bar{y}_U \) and \( t \)

• For a relatively small number \( n \) of sampled primary sampling units, the following confidence intervals are recommended:

\[
\hat{\bar{y}}_{U\text{sys}} \pm t^* \sqrt{\hat{V}(\hat{\bar{y}}_{U\text{sys}})} \\
\hat{t}_{\text{sys}} \pm t^* \sqrt{\hat{V}(\hat{t}_{\text{sys}})} 
\]

where \( t^* \) is the upper \( \alpha/2 \) critical value from the \( t(n-1) \) distribution. Note that the degrees of freedom are based on \( n \), the number of sampled primary sampling units, and not on the total number of secondary sampling units \( nM \).

Systematic Sampling Examples
7.4.4 Comments from W.G. Cochran

- Cochran (from *Sampling Techniques* (1953)) makes the following comments about advantages of systematic sampling:

  Intuitively, systematic sampling seems likely to be more precise than simple random sampling. In effect, it stratifies the population into \([N]\) strata, which consist of the first \([M]\) units, the second \([M]\) units, and so on. We might therefore expect the systematic sample to be about as precise as the corresponding stratified random sample with one unit per stratum. The difference is that with the systematic sample the units all occur at the same relative position in the stratum, whereas with the stratified random sample the position in the stratum is determined separately by randomization within each stratum. The systematic sample is spread more evenly over the population, and this fact has sometimes made systematic sampling considerably more precise than stratified random sampling.

- Cochran also warns us that:

  The performance of systematic sampling relative to that of stratified or simple random sampling is greatly dependent on the properties of the population. There are populations for which systematic sampling is extremely precise and other for which it is less precise than simple random sampling. For some populations and values of \([M]\), \([ \text{var}(\hat{y}_{U,sys}) ]\) may even increase when a larger sample is taken — a startling departure from good behavior. Thus it is difficult to give general advice about the situation in which systematic sampling is to be recommended. A knowledge of the structure of the population is necessary for its most effective use.

- If a population contains a linear trend:

  1. The variances of the estimators from systematic and stratified sampling will be smaller than the variance of the estimator from simple random sampling.

  2. The variance of the estimator from systematic sampling will be larger than the variance of the estimator from stratified sampling. Why? If the starting point of the systematic sample is selected too low or too high, it will be too low or too high across the population of units. Whereas, stratified sampling gives an opportunity for within-stratum errors to cancel.

- Suppose a population has 16 secondary sampling units \((t = 130)\) and is ordered as follows:

```
<table>
<thead>
<tr>
<th>Sampling unit</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>y-value</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
</tbody>
</table>
```

Note there is a linearly increasing trend in the \(y\)-values with the order of the sampling units. Suppose we take a 1-in-4 systematic sample. The following table summarizes the four possible 1-in-4 systematic samples.

```
<table>
<thead>
<tr>
<th>Sampling unit</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>y-values</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>(t_i)</td>
<td>26</td>
<td>104</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\hat{t}_{sys})</td>
<td>30</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 2</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>19</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>(t_i)</td>
<td>35</td>
<td>140</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\hat{t}_{sys})</td>
<td>39</td>
<td>156</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
• If a population has periodic trends, the effectiveness of the systematic sample depends on the relationship between the periodic interval and the systematic sampling interval or pattern. The following idealized curve was given by Cochran to show this. The height of the curve represents the population $y$-value.

  – The $A$ sample points represent the least favorable systematic sample because whenever $M$ is equal to the period, every observation in the systematic sample will be similar so the sample is no more precise than a single observation taken at random from the population.

  – The $B$ sample points represent the most favorable systematic sample because $M$ is equal to a half-period. Every systematic sample has mean equal to the true population mean because successive $y$-value deviations above and below the mean cancel. Thus, the variance of the estimator is zero.

  – For other values of $M$, the sample has varying degrees of effectiveness that depends on the relation between $M$ and the period.

![Idealized curve](image)

7.5 Using a Single Systematic Sample

• Many studies generate data from a systematic sample based on a single randomly selected starting unit (i.e., there is only one randomly selected primary sampling unit).

• When there is only one primary sampling unit, it is possible to get unbiased estimators $\hat{\theta}_{U_{sys}}$ and $\hat{t}_{sys}$ of $\theta_U$ and $t$. It is not possible, however, to get an unbiased estimator of the variances $\hat{V}(\hat{\theta}_{U_{sys}})$ and $\hat{V}(\hat{t}_{sys})$.

• If we can ignore the fact that the $y_{ij}$-values were collected systematically and treat the $M$ secondary sampling units in the single primary sampling unit as a SRS, then the SRS variance estimator would be a reasonable substitute only if the units of the population can reasonably be conceived as being randomly ordered (i.e., there is no systematic pattern in the population such as a linear trend or a periodic pattern).

  – If this assumption is reasonable, then $\hat{V}(\hat{\theta}_{U_{sys}}) \approx \hat{V}(\hat{\theta}_U) = \left( \frac{N-n}{N} \right) \frac{s^2}{n}$

• With natural populations in which nearby units are similar to each other (spatial correlation), this procedure tends to provide overestimates of the variances of $\hat{\theta}_{U_{sys}}$ and $\hat{t}_{sys}$.