in the alphabet is associated with the characteristic of interest. However, systematic sampling is not the same as simple random sampling; it does not have the property that every possible group of n units has the same probability of being the sample. In the preceding example, it is impossible to have students 345 and 346 both appear in the sample. Systematic sampling is technically a form of cluster sampling, as will be discussed in Chapter 5.

Most of the time, a systematic sample gives results comparable to those of an SRS, and SRS methods can be used in the analysis. If the population is in random order, the systematic sample will be much like an SRS. The population itself can be thought of as being mixed. In the quote at the beginning of the chapter, Sorensen reports that President Kennedy used to read a systematic sample of letters written 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, although 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.

2.7 Randomization Theory Results for Simple Random Sampling*[1]

In this section we show that $\bar{y}$ is an unbiased estimator of $\bar{y}_U$: $\bar{y}_U$ is the average of all possible values of $\bar{y}_S$ if we could examine all possible SRSs $\mathcal{S}$ that could be chosen. We also calculate the variance of $\bar{y}$ given in Equation (2.7) and show that the estimator in Equation (2.9) is unbiased over repeated sampling.

[1] An asterisk (*) indicates a section, chapter, or exercise that requires more mathematical background.

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 μ. In the **randomization theory** (also called **design-based**) approach to sampling, the y_i's are considered to be fixed but unknown numbers—any probabilities used arise from the probabilities of selecting units to be in the sample. The randomization theory approach provides a **nonparametric** approach to inference—we need not make any assumptions about the distribution of random variables.

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

$$Z_i = \begin{cases} 1 & \text{if unit } i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\bar{y} = \sum_{i \in \mathcal{S}} \frac{y_i}{n} = \sum_{i=1}^{N} Z_i \frac{y_i}{n}.$$

The Z_i's are the *only* random variables in the above equation 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

$$\pi_i = P(Z_i = 1) = P(\text{select unit } i \text{ in sample}) = \frac{n}{N}. \tag{2.18}$$

The probability in (2.18) 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 Equation (2.18),

$$E[Z_i] = E[Z_i^2] = \frac{n}{N}$$

and

$$E[\bar{y}] = E\left[\sum_{i=1}^{N} Z_i \frac{y_i}{n}\right] = \sum_{i=1}^{N} \frac{n}{N} \frac{y_i}{n} = \sum_{i=1}^{N} \frac{y_i}{N} = \bar{y}_U.$$

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 $i \neq j$,

$$\begin{aligned} E[Z_i Z_j] &= P(Z_i = 1 \text{ and } Z_j = 1) \\ &= P(Z_j = 1 \mid Z_i = 1) P(Z_i = 1) \\ &= \left(\frac{n-1}{N-1}\right)\left(\frac{n}{N}\right). \end{aligned}$$

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 $i \neq j$,

$$\begin{aligned} \text{Cov}(Z_i, Z_j) &= E[Z_i Z_j] - E[Z_i]E[Z_j] \\ &= \frac{n-1}{N-1}\frac{n}{N} - \left(\frac{n}{N}\right)^2 \\ &= -\frac{1}{N-1}\left(1 - \frac{n}{N}\right)\left(\frac{n}{N}\right). \end{aligned}$$

We use the covariance (Cov) of Z_i and Z_j to calculate the variance of $\bar{y}$; see Appendix B for properties of covariances. The negative covariance of Z_i and Z_j is the source of the fpc.

$$\begin{aligned} V(\bar{y}) &= \frac{1}{n^2} V\left(\sum_{i=1}^{N} Z_i y_i\right) \\ &= \frac{1}{n^2} \text{Cov}\left(\sum_{i=1}^{N} Z_i y_i, \sum_{j=1}^{N} Z_j y_j\right) \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(Z_i) + \sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j \,\text{Cov}(Z_i, Z_j)\right] \\ &= \frac{1}{n^2}\left[\frac{n}{N}\left(1 - \frac{n}{N}\right)\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j \frac{1}{N-1}\left(1 - \frac{n}{N}\right)\left(\frac{n}{N}\right)\right] \\ &= \frac{1}{n^2}\frac{n}{N}\left(1 - \frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j\right] \\ &= \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N(N-1)}\left[(N-1)\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2 + \sum_{i=1}^{N} y_i^2\right] \\ &= \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N(N-1)}\left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right] \\ &= \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \end{aligned}$$

To show that the estimator in (2.9) 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 $S^2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/(N-1)$, 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:

$$
\begin{aligned}
E\left[\sum_{i\in\mathcal{S}}(y_i-\bar{y})^2\right] &= E\left[\sum_{i\in\mathcal{S}}\{(y_i-\bar{y}_U)-(\bar{y}-\bar{y}_U)\}^2\right] \\
&= E\left[\sum_{i\in\mathcal{S}}(y_i-\bar{y}_U)^2 - n(\bar{y}-\bar{y}_U)^2\right] \\
&= E\left[\sum_{i=1}^{N} Z_i(y_i-\bar{y}_U)^2\right] - nV(\bar{y}) \\
&= \frac{n}{N}\sum_{i=1}^{N}(y_i-\bar{y}_U)^2 - \left(1-\frac{n}{N}\right)S^2 \\
&= \frac{n(N-1)}{N}S^2 - \frac{N-n}{N}S^2 \\
&= (n-1)S^2.
\end{aligned}
$$

Thus,

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

2.8 A Model for Simple Random Sampling*

Unless you have studied randomization theory in the design of experiments, t proofs in the preceding section probably seemed strange to you. The random variab in randomization theory are not concerned with the responses y_i: They are simp random variables that tell us whether the ith unit is in the sample or not. In a desig based, or randomization theory, approach to sampling inference, the only relationsh between units sampled and units not sampled is that the nonsampled units cou have been sampled had we used a different starting value for the random numb generator.

In Section 2.7 we found properties of the sample mean $\bar{y}$ using randomizatio theory: $y_1, y_2, \ldots, y_N$ were considered to be fixed values, and $\bar{y}$ is unbiased becaus the average of $\bar{y}_S$ for all possible samples $\mathcal{S}$ equals $\bar{y}_U$. The only probabilities use in finding the expected value and variance of $\bar{y}$ are the probabilities that units ar included in the sample.

In your basic statistics class, you learned a different approach to inference. There you had random variables $\{Y_i\}$ that followed some probability distribution, and th actual sample values were realizations of those random variables. Thus you assumed for example, that $Y_1, Y_2, \ldots, Y_n$ were independent and identically distributed fron a normal distribution with mean μ and variance σ^2 and used properties of indepen dent random variables and the normal distribution to find expected values of variou statistics.

We can extend this approach to sampling by thinking of random variables Y_1, $Y_2, \ldots, Y_N$ generated from some model. The actual values for the finite population