Chapter 2

SIMPLE RANDOM SAMPLING

Simple random sampling (SRS) provides a natural starting point for a discussion of probability sampling methods, not because it is widely used, because it is not, but because it is the simplest method, and it underlies many of the more complex methods. As a prelude to defining SRS, we will introduce the notation that the sample size is denoted by n and the finite population size by N. Then, formally defined, SRS is a sampling scheme with the property that any of the possible subsets of n distinct elements from the population of N elements is equally likely to be the chosen sample. This definition implies that every element in the population has the same probability of being selected for the sample, but the definition is more stringent than this. As we will see later, more complex sampling methods are also often equal probability selection methods (*epsem*). However, with such designs, the probabilities of all subsets of the sampled elements of a given size being selected are not all equal, as they are with SRS.

We will illustrate SRS by means of an example. Suppose that a survey is to be conducted in a high school to find out about the students' leisure habits. A list of the school's 1872 students is available, with the list being ordered by the students' identification numbers. These numbers range from 0001 to 1917, with a few gaps in the sequence occurring because some of the students with allocated numbers have since left the school (i.e., they are blanks on the sampling frame). Suppose that an SRS of $n = 250$ is required for the survey. (The choice of n is discussed in Chapter 13.)

One way to draw the required SRS would be by a lottery method. Each student's name or identification number is recorded on one of a set of 1872 identical discs. The discs are placed in an urn, they are thoroughly mixed, and then 250 of them are selected haphazardly. If these operations were executed perfectly, the selected discs would constitute an SRS of 250 students. Although conceptually simple, this method is cumbersome to execute, and it depends on the questionable assumption that the discs have been thoroughly mixed; consequently, it is not used in practice.

Another way of selecting the SRS is by means of a table of random numbers. These tables have been carefully constructed and tested to ensure that in the long run, each digit, each pair of digits, and so on, appears with the same frequency. The volume by RAND (2001) reproduces a widely used random number table that contains a million random digits (originally produced in 1947) and describes how the digits were obtained using Monte

Table 2.1 Random Sampling Numbers

67	28	96	25	68	36	24	72	03	85	49	24
85	86	94	78	32	59	51	82	86	43	73	84
40	10	60	09	05	88	78	44	63	13	58	25
94	55	89	48	90	80	77	80	26	89	87	44
11	63	77	77	23	20	33	62	62	19	29	03

SOURCE: Kendall, M. G. and B. B. Smith, *Tables of Random Sampling Numbers.* Copyright © 1939 by Cambridge University Press.

Carlo methods and the various statistical tests that were performed to check on any departures from randomness. Table 2.1 presents an extract from another table of random numbers, one produced by Kendall and Smith (1939).

Since the student identification numbers comprise four digits, we need to select the random numbers in sets of four. In practice, one should start at some casually chosen point in the table, but here for simplicity, we will start at the top left-hand corner. We will then proceed down the first set of four columns, down the second set of four columns, and so on. Numbers outside the range of the student numbers (0001–1917), and numbers within the range but not associated with a current student, are ignored. Since the first four numbers in the table (6728, 8586, 4010, and 9455) do not yield selections, the first student selected is 1163 (provided this student is still at the school). Continuing through the table, the only other selections from this part of the table are 0588 and 0385. It is already clear that the selection of 250 students in this way is a tedious task, requiring a large selection of random numbers, most of which are nonproductive.

The rejection of so many random numbers can be avoided by associating each student with several random numbers instead of just one; provided that all the students are associated with the same number of random numbers, the sample remains an SRS. Here each student could be associated with 5 four-digit random numbers. A simple scheme is to associate student 0001 also with 2001, 4001, 6001, and 8001; student 0002 also with 2002, 4002, 6002, 8002; and so on through student 1917, who is associated also with 3917, 5917, 7917, and 9917. Then, again starting at the top left-hand corner of the table, the selected students are 6728, i.e., student 0728; 8586, i.e., student 0586; 4010, i.e., student 0010; 9455, i.e., student 1455; 1163, i.e., student 1163; and so on.

These days, random number tables have been largely replaced by computer-generated random numbers. The tables can still sometimes be

useful when the list frame is not computerized but, with computerized lists, the use of a random number generator program makes the selection of a random sample less onerous. There are many random-number generators available, none of which produces truly random numbers, but their pseudo-random numbers are generally adequate for the purpose (although well tested, tables of random numbers are also not perfect). One form of random-number generator creates random numbers from 0 up to, but less than, 1. In this case, one way of selecting a simple random sample is as follows: using the random generator program, assign a different random number to each element in the population, order the population elements by the values of their random numbers, and select the first n elements in the reordered list as the sample.

In drawing the sample using one version of the lottery method or using a table of random numbers, an element could be selected more than once. However, this possibility does not exist with the random-number generation method. With the lottery method, there is a choice of whether or not to replace the disc of an element selected at one draw before the next draw is made. If the disc is not replaced once selected, each element can be selected only once. However, if at each draw the selected disc were replaced in the urn before the next selection is made, elements could be selected more than once. A sample of n discs must contain n distinct elements if sampling is carried out *without replacement*, but the sample may contain fewer than n distinct elements if the sample is drawn *with replacement*. When the sampling procedures described here are conducted with replacement, the sampling method is known as *unrestricted random sampling* or SRS with replacement. When they are conducted without replacement, the method is known just as *simple random sampling* or SRS without replacement. Since sampling without replacement gives more precise estimators than sampling with replacement, we will concentrate on the without-replacement method.

Having selected the SRS of 250 students, we will now assume that the data have been collected from all those sampled (issues of nonresponse are taken up in Chapter 9). The next step is to summarize the individual responses in various ways to provide estimates of characteristics of interest for the population, for instance, the average number of hours of television viewing per day and the proportion of students currently reading a novel. At this point, we need to introduce some notation. Following a common convention in the survey sampling literature, capital letters are used for population values and parameters, and lowercase letters for sample values and estimators. Thus $Y_1, Y_2, Y_3, ..., Y_N$ denote the values of the variable y (e.g., hours of television viewing) for the N elements in the population, and $y_1, y_2, y_3, ..., y_n$ denote the values for the n sampled elements. In general,

the value of variable y for the i-th element in the population is Y_i ($i = 1, 2, ..., N$) and that for the i-th element in the sample is y_i ($i = 1, 2, ..., n$). The population mean of the y-variable is given by

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i,$$

and the sample mean by

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The element variances of the y-variable in the population and in the sample are generally defined in the survey sampling literature as

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (Y_i - \overline{Y})^2,$$

and

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

Sometimes, however, the population element variance is defined with a denominator of N rather than $(N - 1)$, in which case it is denoted by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \overline{Y})^2,$$

that is, $\sigma^2 = (N-1)S^2/N$.

Suppose that we wish to use the data collected in the survey to estimate the mean number of hours of television viewing per day for all students in the school, $\overline{Y}$. This raises the question: How good is the sample mean $\overline{y}$ as an estimator of $\overline{Y}$? This question is unanswerable for a specific estimate $\overline{y}$ from a particular sample; instead, reliance has to be placed on the properties of the estimator on average over repeated applications of the sampling method. Observe here that the term *estimate* is used for a specific value, while *estimator* is used for the rule of procedure used for obtaining the

estimate. In the present example, an estimate of the average number of hours of television viewing may be computed by substituting the values obtained from the sampled students in the estimator $\bar{y} = \Sigma y_i / n$, say $\bar{y} = 2.192$ hours. The theory behind design-based inference provides a means of evaluating estimators but not estimates. The following paragraphs briefly review the design-based theory of statistical inference in the context of SRS.

The design-based properties of sample estimators are derived for a given sample design and form of the estimator. In the present example, suppose that the operations of drawing an SRS of 250 students from the 1872 students and then calculating the sample mean for each sample were carried out an infinite number of times (replacing each sample in the population before drawing the next sample). The resulting set of sample means would have a distribution, known as the *sampling distribution* of the mean. With an SRS sample design, statistical theory shows that the mean of the sampling distribution of the sample mean is the population mean, $\bar{Y}$. In general, if the mean of the individual sample estimates over an infinite number of samples of the given design equals the population parameter being estimated, then the estimator is said to be an *unbiased* estimator of that parameter. Thus, in the case of an SRS, $\bar{y}$ is an unbiased estimator of $\bar{Y}$. Statistical theory also shows that the sampling distribution of the sample mean from a SRS closely approximates the *normal distribution*, provided that the sample size is not too small (an n of 10 or 20 is often sufficient).

Although the sampling distribution of $\bar{y}$ is centered on $\bar{Y}$, any one estimate will differ from $\bar{Y}$; hence, a measure of the variability of the individual estimates around $\bar{Y}$ is needed. A common measure of variability is the standard deviation, the square root of the variance. In this case, the required standard deviation is that of the sample means in the sampling distribution. To avoid confusion with the standard deviation of the element values, standard deviations of sampling distributions are known as *standard errors*. We denote the sample mean of an SRS by $\bar{y}_0$ (with the subscript 0 to indicate SRS), its standard error by $SE(\bar{y}_0)$, and the square of the standard error, the variance of $\bar{y}_0$, by $V(\bar{y}_0)$. For convenience, most sampling error formulas will be presented in terms of variances rather than standard errors. From sampling theory, the variance of a sample mean from an SRS of size n is given by

$$V(\bar{y}_0) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n},$$

or equivalently, and more conveniently, by

$$V(\bar{y}_0) = \left(\frac{N-n}{N}\right)\frac{S^2}{n} = (1-f)\frac{S^2}{n}, \tag{2.1}$$

where $f = n/N$ is the sampling fraction.

These formulas show that $V(\bar{y}_0)$ depends on three factors:

- $(N-n)/(N-1)$ or $(1-f)$, either of which is called the *finite population correction* (fpc)—there is a negligible difference between these terms when N is large;
- n, the sample size; and
- S^2 or σ^2, the alternative versions of the element variance of the y-values in the population.

The fpc term reflects the fact that sampling is conducted without replacement and that the survey population is finite in size, unlike the infinite populations assumed in standard statistical theory. When sampling is conducted with replacement or with an infinite population, there would be no fpc term, in which case Equation (2.1) reduces to the familiar form found in statistical texts, $V(\bar{y}) = \sigma^2/n$. The fpc term indicates the gains of sampling without replacement over sampling with replacement. For samples of size 2 or greater, the fpc term is less than 1, which indicates that $\bar{y}_0$ calculated from an SRS is more *precise*—that is, has a smaller variance—than $\bar{y}$ calculated from an unrestricted sample of the same size. In many practical situations, populations are large and, even though the samples may also be large, the sampling fractions are small. In such situations, the difference between sampling with and without replacement is unimportant because, even if the sample were drawn with replacement, there would be little chance that an element would be selected more than once. This argument can also be expressed in terms of the fpc term. If the sampling fraction (f) is say 1/10, the fpc term is 0.9, and its effect on the standard error is as a multiplier $\sqrt{1-f} = 0.95$; if $f = 1/20$, $(1-f) = 0.95$, and $\sqrt{1-f} = 0.97$. Thus, if the sampling fraction is small, the fpc term is close to 1, and it has a negligible effect on the standard error. The fpc term is often neglected (i.e., treated as 1) when the sampling fraction is less than 1 in 20, or even less than 1 in 10.

The second factor in the formula for $V(\bar{y}_0)$ is the sample size, n. As is intuitively obvious, the larger the sample, the smaller is $V(\bar{y}_0)$. What is perhaps less obvious is the fact that, for large populations, it is the sample size rather than the sampling fraction that is dominant in determining the precision of survey results. As a consequence, estimates obtained from a sample of size 2000 drawn from a country with a population of 300 million

are about as precise as those obtained from a sample of the same size drawn from a small city of 40,000 (assuming the element variances in the two populations are the same). It also follows from this line of argument that the gains from sampling are greatest with large populations. Indeed, for very small populations, the gains from sampling may not be worthwhile, even though the fpc term has an appreciable effect in such cases. For example, it may be more convenient to take all students in a school of 200, rather than sample 175 of them.

The third factor in the formula for $V(\overline{y}_0)$ is the element variance of the y-values in the population, either σ^2 or S^2. Clearly, if all the students watch approximately the same amount of television, the mean of any sample will be close to the population mean. However, if the students differ greatly in their viewing habits, there is a risk that the sample mean will differ considerably from the population mean. Note that σ^2 and S^2 are unknown population parameters. An estimate of the population element variance is needed to estimate $V(\overline{y}_0)$. The advantage of using Equation (2.1) for $V(\overline{y}_0)$, expressed in terms of S^2, is that the familiar sample estimator $s^2 = \Sigma(y_i - \overline{y})^2/(n-1)$ is an unbiased estimator for S^2 (but not for σ^2). Thus $V(\overline{y}_0)$ and $SE(\overline{y}_0)$ may be simply estimated by

$$v(\overline{y}_0) = (1-f)s^2/n \tag{2.2}$$

and

$$se(\overline{y}_0) = \sqrt{(1-f)s^2/n}, \tag{2.3}$$

with lowercase letters for $v(\overline{y}_0)$ and $se(\overline{y}_0)$ to indicate sample estimators.

Having estimated the standard error, a confidence interval can be calculated for the population mean. With a large SRS sample, the sampling error arising from replacing S^2 by s^2 can be ignored. Then the 95% confidence interval for $\overline{Y}$ is $\overline{y}_0 \pm 1.96 se(\overline{y}_0)$, where the multiplier 1.96 is taken from a table of the normal distribution (95% of the normal distribution falls within 1.96 standard deviations around the distribution's mean). As an illustration, suppose that the mean hours watching television per day for the 250 sampled students is $\overline{y}_0 = 2.192$ hours, with an element variance of $s^2 = 1.008$. A 95% confidence interval for $\overline{Y}$ is then given by

$$2.192 \pm 1.96\sqrt{\left(1 - \frac{250}{1872}\right)\frac{1.008}{250}} = 2.192 \pm 0.116.$$

That is, we are 95% confident that the interval from 2.076 to 2.308 contains the population mean.

The use of the normal distribution for calculating confidence intervals applies for large samples in which the sample element variance s^2 estimates the population element variance S^2 with high precision. When the sample size is less than, say, 30, the normal distribution should be replaced by Student's *t* distribution, thus widening the interval to take account of the sampling error in s^2 (see, for example, Dietz & Kalof, 2009).

The proportion (or percentage) of the population with a particular attribute, for instance the proportion of students currently reading a novel, is a parameter of common analytic interest. Results for a proportion follow directly from those for a mean, since a proportion is just a special case of a mean that is obtained by setting $Y_i = 1$ if the *i*-th element has the attribute and $Y_i = 0$ if not. Then $\overline{Y} = \Sigma Y_i / N$ is simply P, the population proportion with the attribute, and the sample mean $\overline{y}$ is the sample proportion p. Thus, in general, the theoretical results obtained for a sample mean apply also for a proportion. In the case of SRS, since $\overline{y}_0$ is unbiased for $\overline{Y}$, it follows that p_0 is unbiased for P. The standard error and variance formulas given above for $\overline{y}_0$ can also be applied to p_0. However, since the *y*-variable takes values of only 0 or 1 for a proportion, the formulas for S^2 and s^2 can be simplified to $NPQ/(N-1)$ and $np_0q_0/(n-1)$, respectively, where $Q = 1 - P$ and $q_0 = 1 - p_0$. Using these simplifications,

$$V(p_0) = (1-f)\frac{NPQ}{(N-1)n} \tag{2.4}$$

and

$$v(p_0) = (1-f)\frac{p_0 q_0}{(n-1)}. \tag{2.5}$$

If the fpc term can be neglected, and if n is large, $v(p_0)$ reduces as an approximation to the well-known formula p_0q_0/n. These formulas also apply with P and p_0 expressed as percentages, with the modifications that $Q = 100 - P$ and $q_0 = 100 - p_0$.

As an illustration, suppose that 165 of the 250 sampled students were reading a novel, i.e., $p_0 = 66.0\%$. Using the same approach as for the sample mean, a 95% confidence interval for P is then

$$66.0 \pm 1.96\sqrt{\left(1 - \frac{250}{1872}\right)\frac{66.0 \times 34.0}{249}} = 66.0\% \pm 5.5\%,$$

that is, we are 95% confident that the interval 60.5% to 71.5% contains the population percentage.

Although widely used, this form of confidence interval for a proportion—known as a Wald interval—has the limitation that the standard error estimate for a sample proportion is a function of that proportion. As a result, the coverage properties of the Wald interval can be seriously in error when the proportion is outside the middle of the range from 0 to 1. A number of alternative forms of the confidence interval for a proportion have been proposed in an attempt to address this problem (see Brown, Cai, & DasGupta, 2001; Dean & Pagano, 2015; and Franco, Little, Louis, & Slud, 2019, for evaluations of some alternatives). In this book, we apply the Wald interval, although in practice, the Wilson interval may better reflect the coverage properties of the interval when the proportion is ≤ 0.2 or ≥ 0.8.

The preceding discussion reviews the steps involved in estimating a population mean or proportion from an SRS and calculating an associated confidence interval. The same approach can also be used for the estimation of other population parameters. The only feature that distinguishes design-based inference with an SRS sample from model-based inference is the inclusion of the fpc term, which can often be ignored. However, as shown in later chapters, model-based formulas should not be used to produce design-based inferences with other sample designs.