3
Simple Random Sampling

3.1 INTRODUCTION

Everyone mentions simple random sampling, but few use this method for population-based surveys. Rapid surveys are no exception, since they too use a more complex sampling scheme. So why should we be concerned with simple random sampling? The main reason is to learn the theory of sampling. Simple random sampling is the basic selection process of sampling and is easiest to understand.

If everyone in a population could be included in a survey, the analysis featured in this book would be very simple. The average value for equal interval and binomial variables, respectively, could easily be derived using Formulas 2.1 and 2.3 in Chapter 2. Instead of estimating the two forms of average values in the population, they would be measuring directly. Of course, when measuring everyone in a population, the true value is known; thus there is no need for confidence intervals. After all the purpose of the confidence interval is to tell how certain the author is that a presented interval brackets the true value in the population. With everyone measured, the true value would be known, unless of course there were measurement or calculation errors.

When the true value in a population is estimated with a sample of persons, things get more complicated. Rather then just the mean or proportion, we need to derive the standard error for the variable of interest, used to construct a confidence interval. This chapter will focus on simple random sampling or persons or households, done both with and without replacement, and present how to derive the standard error for equal interval variables, binomial variables, and ratios of two variables. The latter, as described earlier, is commonly used in rapid surveys and is termed a *ratio estimator*. What appears to be a proportion, may actually be a ratio estimator, with its own formula for the mean and standard error.

3.1.1 Random sampling

Subjects in the population are sampled by a random process, using either a random number generator or a random number table, so that each person remaining in the population has the same probability of being selected for the sample. The process for selecting a random sample is shown in Figure 3-1.

Figure 3-1

The population to be sampled is comprised of nine units, listed in consecutive order from one to nine. The intent is to randomly sample three of the nine units. To do so, three random numbers need to be selected from a random number table, as found in most statistics texts and presented in Figure 3-2. The random number table consists of six columns of two-digit non-repeatable numbers listed in random order. The intent is to sample three numbers between 1 and 9, the total number in the population. Starting at the top of column A and reading down, two numbers are selected, 2 and 5. In column B there are no numbers between 1 and 9. In column C the first random number in the appropriate interval is 8. Thus in our example, the randomly selected numbers are 2, 5 and 8 used to randomly sample the subjects in Figure 3-1. Since the random numbers are mutually exclusive (i.e., there are no duplicates), each person with the illustrated method is only sampled once. As described later in this chapter, such selection is sampling *without replacement*.

Figure 3-2

Random sampling assumes that the units to be sampled are included in a list, also termed a sampling *frame*. This list should be numbered in sequential order from one to the total number of units in the population. Because it may be time-consuming and very expensive to make a list of the population, rapid surveys feature a more complex sampling strategy that does not require a complete listing. Details of this more complex strategy are presented in Chapters 4 and 5. Here, however, every member of the population to be sampled is listed.

3.1.2 Nine drug addicts

A population of nine drug addicts is featured to explain the concepts of simple random sampling. All nine addicts have injected heroin into their veins many times during the past weeks, and have often shared needles and injection equipment with colleagues. Three of the nine addicts are now infected with the human immunodeficiency virus (HIV). To be derived are the proportion who are HIV infected (a binomial variable), the mean number of intravenous injections (IV) and shared IV injections during the past two weeks (both equal interval variables), and the proportion of total IV injections that were shared with other addicts. This latter proportion is a ratio of two variables and, as you will learn, is termed a *ratio estimator*.

Figure 3-3

The total population of nine drug addicts is seen in Figure 3-3. Names of the nine male addicts are listed below each figure. The three who are infected with HIV are shown as cross-hatched figures. Each has intravenously injected a narcotic drug eight or more times during the past two weeks. The number of injections is shown in the white box at the midpoint of each addict. With one exception, some of the intravenous injections were shared with other addicts; the exact number is shown in Figure 3-3 as a white number in a black circle.

Our intention is to sample three addicts from the population of nine, assuming that the entire population cannot be studied. To provide an unbiased view of the population, the sample mean

should on average equal the population mean, and the sample variance should on average equal the population variance, corrected for the number of people in the sample. When this occurs, we can use various statistical measures to comment about the truthfulness of the sample findings. To illustrate this process, we start with the end objective, namely the assessment of the population mean and variance.

Population Mean. For total intravenous drug injections, the mean in the population is derived using Formula 3.1

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{3.1}$$

where X_i is the total injections for each of the i addicts in the population and N is the total number of addicts. Thus, the mean number of intravenous drug injections in the population shown in Figure 3-3 is

$$\bar{X} = \frac{10+8+12+9+11+11+9+11+10}{9} = \frac{91}{9} = 10.1$$

or 10.1 intravenous drug injections per addict.

Population Variance. Formula 3.2 is used to calculate the variance for the number of intravenous drug injections in the population of nine drug addicts.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N} \tag{3.2}$$

where σ^2 is the Greek symbol for the population variance, X_i and N are as defined in Formula 3.1 and $\bar{X}$ is the mean number of intravenous drug injections per addict in the population. Using Formula 3.2, the variance in the population is

$$\sigma^2 = \frac{(10-10.1)^2 + (8-10.1)^2 \ldots (11-10.1)^2 + (10-10.1)^2}{9} = 1.43$$

Sample Mean. Since the intent is to make a statement about the total population of nine addicts, a sample of three addicts will be drawn, and their measurements will be used to represent the group.