4
Equal Probability of Selection


4.1 INTRODUCTION

When is a person treated as a person? For rapid surveys, the answer is when the sample is self-weighted. Statistical calculations are easier when each person in a population has the same chance of being selected as any other person. On occasion, we may draw a sample that is not self-weighted. When this occurs, we need to add numeric weights to our estimates of the average value and standard error, so that the sample again gives a factual view of the population. The use of such weights will be described in Chapter 6. For now, however, I will focus on self-weighted sampling.

4.1.1 Sample of ten from population of 120

In Chapter 2 we learned of simple random sampling with a population of nine addicts and samples of three addicts. We observed that with such a small population there are only 84 possible samples that could be drawn of three addicts without replacement. I concluded the chapter with an example of a survey of smoking patterns, with a sample of 30 households from 1,000 households and a sample of 90 persons from a population of 3,000 persons. We discovered that the number of possible samples is greater than we could easily count, or even that a computer could count. As a result, we had to draw samples from a huge list of all possible samples so as to create and view 100 confidence intervals. Here I will present still another sample but this time of ten people from a population of 120. I will use this example to illustrate what it means to be a self-weighted sample and for each person to have the same probability of being selected.
Assume there is a population of 120 persons from which we are drawing a random sample of 10 persons (see Figure 4-1). All of the persons were sampled without replacement so that the 10 persons in the sample are all different from one another. There are many possible combinations of 10 people that could be selected from the population. Using Formula 3.2, we can calculate the exact number of possible samples as:

\[
\frac{120!}{10! (120-10)!}
\]

Extending the formula, the calculations are...

\[
\frac{120 \times 119 \times 118 \times 117 \times 116 \times 115 \times 114 \times 113 \times 112 \times 111 \times 110!}{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 \times 110!}
\]

The term 110! cancels in the numerator and denominator,

\[
\frac{120 \times 119 \times 118 \times 117 \times 116 \times 115 \times 114 \times 113 \times 112 \times 111}{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}
\]

resulting in an answer of 116,068,178,638,776 possible samples. If we had much paper and endless time, we could list all 116 trillion possible samples, as represented in Figure 4-2. Keep in mind that the one sample we actually selected shown in the lower left of Figure 4-2 is merely one of the 116 trillion samples we could have selected. While our survey of 10 persons may be viewed as a random sample of 10 individuals from a list of 120 people, you could also consider it a random sample of a group of 10 persons from a list of 116 trillion possible samples. If the sampling method is self-weighted and free of bias, each of these 116 trillion samples has the same probability of being selected. By chance alone, we selected sample number 100,000,000,000,000.

The data in each of the possible samples can be used to derive a proportion or a mean, depending on whether the variable is binomial or equal interval (see Figure 4-3). The mean and proportion of the sample we selected number 100,000,000,000,000 is shown in the lower left section of Figure 4-3, listed among the 116,068,178,638,775 other possible samples. If the sampling method is self-weighted and unbiased, each of the proportions or means from the trillions of samples will have...
the same probability of being selected and we will be able to use the statistical formulae cited in Chapter 3 (and presented in more detail in Chapter 5) to calculate the standard error and confidence interval. If the sample is biased, some of the possible samples may have no chance of being selected. If the persons are not self-weighted, some of the possibles samples will be more likely to be selected than others. Different from some forms of bias, however, we can correct for unequal probabilities of selection with numeric weights that we attach to each sampled person. Such weighting is common in advanced surveys but is not necessary with most types of rapid survey. I will, however, describe weights in Chapter 6, as a useful procedure for combining rapid surveys from several regions of a country, or survey strata of different groups identified by race, sex, education and the like.

4.2 EPSEM SAMPLING

Sampling which results in each person having the same chance of being selected is termed equal probability of selection method (EPSEM) sampling. For sampling done in one stage, such as simple random sampling described in Chapter 2, each person should have the same probability of being selected at the beginning of the stage. If the sample is drawn in two stages, as will be described in this chapter, the probability of being selected may be different at the beginning of stage one than in advance of stage two. EPSEM sampling does not mean that every unit at every stage has the same probability of being selected. Instead it means that every unit that is actually included in the sample had the same probability of being selected and we will be able to use the statistical formulae cited in Chapter 3 (and presented in more detail in Chapter 5) to calculate the standard error and confidence interval. If the sample is biased, some of the possible samples may have no chance of being selected. If the persons are not self-weighted, some of the possibles samples will be more likely to be selected than others. Different from some forms of bias, however, we can correct for unequal probabilities of selection with numeric weights that we attach to each sampled person. Such weighting is common in advanced surveys but is not necessary with most types of rapid survey. I will, however, describe weights in Chapter 6, as a useful procedure for combining rapid surveys from several regions of a country, or survey strata of different groups identified by race, sex, education and the like.
selected in advance of sampling. No one is left out and no one is favored.

### 4.2.1 One-stage sampling

Typical of EPSEM surveys is a simple random sample of women selected in one stage from a population of 2,000. As observed in Figure 4-4, 2,000 women comprise a population. That is, the population is people. Before being surveyed, they must all be included in a larger list termed a frame, with sequential identification numbers starting at one and ending with the total number in the population. In our example the identification numbers go from 1 to 2,000. Using random numbers between 1 and 2,000, ten women are sampled from the frame. Thus the sampling unit is a person. If the sample is self-weighted, each woman has a probability of being selected of 1/2,000 or 0.0005.

The concept is the same when sampling households (see Figure 4-5). Rather than a population of persons, Figure 4-5 shows a population of 1,000 households. To do a simple one-stage random sample, each household is included in a frame with a sequential identification number from one to 1,000. If the sample is self-weighted, each household has the same probability of being selected, shown in Figure 4-5 as 1/1,000 or 0.001. Notice that the sampling unit is households, not people and that the sampled households are selected in one stage from the frame.

The probability of being selected for a sample increases with the number of units that are sampled. Assume that you are sampling 20 units (either people or households) from a population of 10,000 units. Each person has a probability of 1/10,000 or 0.0001.

From basic probability theory we know that for a sample of two units, the probability that a unit is selected at either drawing one or two is defined as in Formula 4.1

\[
P(\text{one or two}) = P(\text{one}) + P(\text{two})
\]  

(4.1)

The selection probability for a sampling unit when there is multiple drawings is the addition of the individual probabilities. Returning to our example of a sample of 20 units from a population of 10,000 units, the selection probability for entering the sample is...

\[
P(\text{one or two or... 19 or 20}) = P(\text{one}) + P(\text{two}) + \ldots P(19) + P(20)
\]
or 20/10,000, as shown in Figure 4-6. To be unbiased, the probability of any one unit being selected should be the same as any other unit. That is, the selection probability for each unit should be 20/10,000 or 0.002. If the selection probabilities are different, the sample units can be mathematically weighted so that each sample unit again represents the same number of units in the population. Such weighting schemes, however, are more complicated and should be avoided when doing rapid surveys. The concept of weighting will be presented further in Chapter 6.

4.2.2 Two-stage sampling

Persons or households can also be selected in two stages. The first stage might be a sample of city blocks or districts, or villages in rural areas. The second stage is a sample of people or households drawn from those units selected in the first stage. An example of the first stage selection process is shown in Figure 4-7 for a small town with 65 blocks. The blocks are listed with sequential numbers from 1 to 65. At the first stage, a sample of nine blocks is drawn from the population of 65 blocks. Notice that time and effort must be spent in listing the blocks at the first stage. There is no easy way to do this step other than to count the blocks.

Only those blocks that are selected in Figure 4-7 at the first stage are included in the second stage of the sampling process. Yet before further sampling can be done, we again need to create a frame (or list) of persons or households in each sampled block. That is, a sampling frame must be created for blocks number 3, 12, 21, 25, 35, 38, 44, 53 and 60, but not for the remaining 56 blocks. If we had done a one-stage simple random sample, we would have to have listed every sampling unit in the population rather than just those in the nine selected blocks. This would have been much more time consuming and is the main reason why simple random sampling is rarely done. The field work is too costly, unless of course there are volunteers available who will do
quality work for little or no pay. For most people, however, two-stage sampling is preferable.

For the second stage, only those blocks to be sampled are included. Thus in our example, a new frame is created with nine blocks (see Figure 4-8). Each of the nine blocks did not have the same chance of being selected. Instead, the blocks at the first stage were selected with probability proportionate to their size (PPS). Using this method, larger blocks are more likely to be included in our sample than smaller blocks. Further details on this selection procedure, featured in rapid surveys, are presented in Section 4.3.

Within the nine selected blocks, all persons or households are listed. A simple random sample is then done in each cluster of the listed units and the data are combined into a single estimate for the survey as a whole. Figure 4-9 illustrates the two-stage sampling process for a survey of women to assess their choice of contraceptive methods for family planning. The 65 city blocks are listed in the first stage frame. Nine of the blocks are sampled from the frame with probability proportionate to size. These nine blocks lead to the creation of another frame, consisting of nine lists of eligible women living in the selected blocks. Once each list is made, seven women are selected from the list by simple random sampling. The total sample consists of 63 women, too small for a rapid survey but large enough for an example.

Rather than persons, the second stage of the sampling process is often households (see Figure 4-10). The sampling process is identical except that the second stage frame consists of a listing of households in each block rather than women. The size of the second survey is 36 households, far smaller than the first survey. But is it? As you will learn later, in household surveys we are generally not interested in characteristics of the household. Instead we want to know more about the occupants of the households. If we collected information on each persons in the 36 households, and there was an average of four persons per household, the survey would consist of 144 persons, compared to 63
with the family planning survey of women.

Conversely, if we only wanted information on women, aged 20 through 44 years, in the households, the number of participants in our sampled households might be much smaller than 63.

### 4.2.3 Two-stage cluster sampling

While the concepts of self-weighting and equal probability of selection are easy to read about, they are sometimes overlooked when planning a survey. In this section I will present four sampling strategies for two-stage cluster surveys, three of which are EPSEM samples and one which is not. Rather than city blocks, households or people, I will use circles and squares to represent sampling units. The population to be sampled usually has some organizational characteristics that links groups of people together at some time of day or year. The group may be defined as living in a common census tract or enumeration district, or in going to the same school, or being eligible to attend the same health clinic. The term cluster is used by sampling specialists to describe units in the population that contain the sampling units of interest. A cluster is formally defined as a natural grouping within a population. This grouping may be a city block, a census tract or enumeration district, a health district, a school district, or a community such as a village, hamlet, or town. For rapid surveys, clusters are usually geographic areas with political or social boundaries. The cluster may vary in size, although on rare occasions the clusters may all be the same size.

To illustrate how procedures at two steps of the sampling process can bias a survey, I will use the same example presented earlier in Figure 4-6. There we sampled 20 units from a population or 10,000 units and the selection probability was 0.002. We will again draw a sample of 20 from 10,000 but do so in two stages to ease the field work. If the sampling scheme is unbiased, each unit in the population should have the same probability of appearing in the sample. In our example, that selection probability is 0.002. I have divided the population into 200 clusters or groupings of units. In Figure 4-11, I show what occurs when the clusters vary in size, with some being as small as 10 units and others being as large as 100 units. At the first stage, four clusters are selected by simple random sampling. Thus each cluster has the same probability of being selected: 1/200 or 0.002. At the second stage, I select an equal fraction of units from each cluster drawn at the first stage, again by simple random sampling. As a result, 10 units would be selected from the larger cluster of 100 units, while one unit would be drawn from the small cluster of 10 unit. How does this two-part selection scheme effect the selection
Figure 4-11. Equal probability of selection with simple random sampling of unequal sized-clusters at first stage and simple random sampling of equal fraction at second stage.

probabilities? To answer this question we need to again use probability theory. Basic probability theory states that for a sampling scheme with two stages, the probability of having a unit selected at both stages is defined as in Formula 4.2

$$P(\text{Stage one and Stage two}) = P(\text{Stage one}) \times P(\text{Stage two})$$

(4.2)

If a unit happens to be in a cluster of 100 units, the probability of being selected with the described sampling scheme is 1/200 at the first stage and 10/100 at the at the second stage for a selection probability of...

$$P(\text{Stage one and Stage two}) = \frac{1}{200} \times \frac{10}{100}$$

or 0.002. This selection probability is the same for all 20 units included in the sample, as it was in Figure 4-6. Hence if clusters vary in size, an unbiased method of sampling is to draw clusters at the first stage by simple random sampling, followed by a simple random sample of a constant fraction of units in the selected clusters at the second stage.

For a second example, assume that all clusters are the same size (see Figure 4-12). Here we draw four clusters at the first stage by simple random sampling. At the second stage, however, we draw an equal number of units in the selected clusters, again by simple random sampling. The selection probabilities for all units in the sample remains at 0.002, indicating that this second sampling scheme also does not favor one unit over another.