Chapter 5

Fundamentals of Applied Sampling

Thomas Piazza

5.1 The Basic Idea of Sampling

Survey sampling is really quite remarkable. In research we often want to know certain characteristics of a large population, but we are almost never able to do a complete census of it. So we draw a sample—a subset of the population—and conduct research on that relatively small subset. Then we generalize the results, with an allowance for sampling error, to the entire population from which the sample was selected. How can this be justified?

The capacity to generalize sample results to an entire population is not inherent in just any sample. If we interview people in a "convenience" sample—those passing by on the street, for example—we cannot be confident that a census of the population would yield similar results. To have confidence in generalizing sample results to the whole population requires a "probability sample" of the population. This chapter presents a relatively non-technical explanation of how to draw a probability sample.

Key Principles of Probability Sampling

When planning to draw a sample, we must do several basic things:

1. **Define carefully the population to be surveyed.** Do we want to generalize the sample result to a particular city? Or to an entire nation? Or to members of a professional group or some other organization? It is important to be clear about our intentions. Often it may not be realistic to attempt to select a survey sample from the whole population we ideally would like to study. In that case it is useful to distinguish between the entire population of interest (e.g., all adults in the U.S.) and the population we will actually attempt to survey (e.g., adults living in households in the continental U.S., with a landline telephone in the home). The entire population of interest is often referred to as the "target population," and the

more limited population actually to be surveyed is often referred to as the "survey population."[1]

2. **Determine how to access the survey population (the sampling frame).** A well-defined population is only the starting point. To draw a sample from it, we need to define a "sampling frame" that makes that population concrete. Without a good frame, we cannot select a good sample. If some persons or organizations in the survey population are not in the frame, they cannot be selected. Assembling a sampling frame is often the most difficult part of sampling. For example, the survey population may be physicians in a certain state. This may seem well-defined, but how will we reach them? Is there a list or directory available to us, perhaps from some medical association? How complete is it?
3. **Draw a sample by some random process.** We must use a random sampling method, in order to obtain results that represent the survey population within a calculable margin of error. Selecting a few convenient persons or organizations can be useful in qualitative research like focus groups, in-depth interviews, or preliminary studies for pre-testing questionnaires, but it cannot serve as the basis for estimating characteristics of the population. Only random sampling allows generalization of sample results to the whole population and construction of confidence intervals around each result.
4. **Know the probability (at least in relative terms) of selecting each element of the population into the sample.** Some random sampling schemes include certain population elements (e.g., persons or organizations) at a higher rate than others. For example, we might select 5% of the population in one region but only 1% in other regions. Knowing the relative probabilities of selection for different elements allows the construction of weights that enable us to analyze all parts of a sample together.

The remainder of this chapter elaborates on and illustrates these principles of probability sampling. The next two sections cover basic methods for sampling at random

[1] This is the terminology introduced by Kish (1965, p. 7) and used by Groves et al. (2009, pp.69-70) and by Kalton (1983, pp. 6-7). This terminology is also used, in a slightly more complicated way, by Frankel (this

from a sampling frame. We proceed to more complicated designs in the sections that follow.

5.2 The Sampling Frame

Developing the frame is the crucial first step in designing a sample. Care must be exercised in constructing the frame and understanding its limitations. We will refer to the frame as a list, which is the simplest type of frame. However, a list may not always be available, and the frame may instead be a *procedure* (such as the generation of random telephone numbers) that allows us to access the members of the survey population. But the same principles apply to every type of frame.

Assemble or identify the list from which the sample will be drawn

Once we have defined the survey population – that is, the persons or organizations we want to survey—how do we find them? Is there a good list? Or one that is "good enough"? Lists are rarely perfect: common problems are omissions, duplications, and inclusion of ineligible elements.

Sometimes information on population elements is found in more than one file, and we must construct a comprehensive list before we can proceed. In drawing a sample of schools, for instance, information on the geographic location of the schools might be in one file, and that on academic performance scores in another. In principle, a sampling frame would simply merge the two files. In practice this may be complicated, if for example the two files use different school identification codes, requiring a "crosswalk" file linking the corresponding codes for a given school in the different files.

Dealing with incomplete lists

An incomplete list leads to **non-coverage error** – that is, a sample that does not cover the whole survey population. If the proportion of population elements missing from the list is small, perhaps 5% or less, we might not worry. Sampling from such a list

volume).

could bias[2] results only slightly. Problems arise when the proportion missing is quite large.

If an available list is incomplete, it is sometimes possible to improve it by obtaining more information. Perhaps a second list can be combined with the initial one. If resources to improve the list are not available, and if it is our only practical alternative, we might redefine the survey population to fit the available list. Suppose we initially hoped to draw a sample of all physicians in a state, but only have access to a list of those in the medical association. That frame omits those physicians who are not members of the association. If we cannot add non-members to that frame, we should make it clear that our survey population includes only those physicians who are members of the medical association. We might justify making inferences from such a sample to the entire population of physicians (the target population) by arguing that non-member physicians are not very different from those on the list in regard to the variables to be measured. But unless we have data to back that up, such arguments are conjectures resting on substantive grounds – not statistical ones.

Duplicates on lists

Ideally a list includes every member of the survey population – but only once. Some elements on a list may be duplicates, especially if a list was compiled from different sources. If persons or organizations appear on a list more than once, they could be selected more than once. Of course, if we select the same element twice, we will eventually notice and adjust for that. The more serious problem arises if we do not realize that an element selected only once had duplicate entries on the frame. An element that appears twice on a list has double the chance of being sampled compared to an element appearing only once, so unrecognized duplication could bias the results. Such differences in selection probabilities should be either eliminated or somehow taken into account (usually by weighting) when calculating statistics that will be generalized to the survey population.

[2] The term “bias” refers to an error in our results that is not due to chance. It is due to some defect in our sampling frame or our procedures.

The most straightforward approach is to eliminate duplicate listings from a frame before drawing a sample. Lists available as computer files can be sorted on any field that uniquely identifies elements—such as a person's or organization's name, address, telephone number, or identification code. Duplicate records should sort together, making it easier to identify and eliminate them. Some duplicates will not be so easily isolated and eliminated, though, possibly because of differences in spelling, or recordkeeping errors.

Alternately, we can check for duplicates after elements are selected. A simple rule is to accept an element into the sample only when its *first* listing on the frame is selected (Kish, 1965, p. 58). This requires that we verify that every selected element is a first listing, by examining the elements that precede the position of that selection on the list. Selections of second or later listings are treated as ineligible entries (discussed next). This procedure can be extended to cover multiple lists. We predefine a certain ordering of the lists, and after selecting an element we check to see that it was not listed earlier on the current list or on the list(s) preceding the one from which the selection was made. This procedure requires that we check only the *selected* elements for duplication (rather than *all* elements on the frame), and that we check only the part of the list(s) preceding each selection.

Ineligible elements

Ineligible elements on a list present problems opposite to those posed by an incomplete list. Ineligible entries are elements that are outside the defined survey population. For example, a list of schools may contain both grade schools and high schools, but the survey population may consist only of high schools. Lists are often out of date, so they can contain ineligible elements—like schools that have closed, or persons who have died.

It is best to delete ineligible elements that do not fit study criteria, if they are easily identified. Nevertheless, ineligible records remaining on the frame do not pose major problems. If a selected record is determined to be ineligible, we simply discard it. One should **not** compensate by, for example, selecting the element on the frame that follows an ineligible element. Such a rule could bias the sample results, because elements immediately following ineligible ones would have higher selection probabilities – their own probability plus that of the immediately preceding ineligible element(s).

When a list includes ineligible entries, we must ensure that the sample includes enough usable selections by anticipating the ineligibility rate and sampling additional elements. If the target sample size is 500, for example, and we expect that 20% of the elements on the frame are ineligible, selecting 500 elements would leave only 400 usable selections. To end up with 500, we should select 500/(1-0.20)=625. If we anticipate further that only 70% of the eligible selected elements (persons or organizations) will agree to participate in the survey, we should increase the sample size even further to 625/0.70 = 893.

Indeed, once we decide on a certain target number of completed interviews, it is usually necessary to make many more than that number of selections, to compensate for anticipated losses due to ineligibles, duplicates, refusals, language problems, and other issues. Such adjustments in sample selection plans are an important part of sampling work.

5.3 Basic Methods for Random Sampling from Lists

Selecting persons, organizations or other elements from a list is the simplest and most straightforward sampling method. It illustrates the main points in sampling and provides groundwork for more complex methods. Variations on the basic theme exist, however, even for this simplest sample selection method.

Once the frame has been assembled, we can draw one or more samples. Three commonly used sampling methods are simple random sampling, systematic sampling, and selection with probability proportional to size.

5.3.1 Simple Random Sampling

Simple random sampling (SRS) is the standard basic method of sampling. With SRS, each element on the list has the same selection probability, and selections are made independently of one another. SRS serves as a baseline against which other methods are evaluated.

Selection can be carried out either “with replacement” or “without replacement.” To understand the terminology, think of selecting little numbered balls from a big jar. If we put a ball back in the jar after selecting it, we could select the same ball more than once. If we do not replace selected balls, we cannot select the same ball more than once. A valid random sample can be drawn either way. The statistical theory of random sampling is a little simpler if sampling is done *with* replacement. In practice, however, we almost always prefer *not* to select the same person or organization more than once, and therefore we usually sample *without* replacement.

Figure 5.1 illustrates a very simple procedure for drawing simple random samples. Suppose we would like to select 2 of the 10 elements in Figure 5.1 at random. We could generate some independent random numbers between 1 and 10 using a spreadsheet, a computer program, or a table of random numbers. In this example we generated (in order) 8, 4, 7, 6, and 6. The first random number selects element #8 on the list, and the second selects element #4.

(Figure 5.1 about here)

The element numbers could refer to the sequential position of elements on the list, or to another unique identifier for each element, so that each random number refers to no more than one element. If the element numbering system has gaps, some random numbers might not correspond to any element. In that case, we simply discard such a random number and move on to the next one.

In Figure 5.1, we generated more than two random numbers even though we wanted only two selections, because we planned to select elements without replacement. Since random numbers are usually generated independently, some could be duplicates. (Indeed, the fourth and the fifth random numbers are both 6.) If a random number is the same as an earlier one, we discard it and move on to the next unique one.

Many lists used as sampling frames are available as computer files. In such cases we can use a spreadsheet or a statistical program such as SPSS, SAS, or Stata to select a simple random sample.

5.3.2 Systematic Random Sampling

Systematic sampling selects elements from a list by using a fixed selection interval, calculated by dividing the number of elements on the list by the desired number of selections. Randomness is introduced by choosing a random number within the first interval to make the first selection. To make subsequent selections, the interval is added successively to the preceding selection number.

For example, to select 20 elements from a list of 100, we use an interval of 100/20 = 5, and we select every 5th element. To begin, we would take a random number between 1 and 5, say 3. Then we would select elements 3, 8, 13, 18, and so on up to 98. The random number should be obtained from a table of random numbers or generated by a computer program, not a number we happened to think of "at random." Notice in this example that there are only five distinct samples of elements that can be selected, corresponding to the five possible random starts between 1 and 5. This simplicity makes the method easy to use, but it has to be used with some care.

Systematic selection is used for many kinds of lists, but it is especially convenient for sampling from lists that are not computerized and when records are not numbered sequentially. One only has to estimate the number of entries on the list, calculate the interval that will produce the desired sample size, generate a random start, and then just count off the selections.

Systematic selection never draws the same element more than once (unless a list has duplicates or occasionally when sampling is done with probability proportional to size, to be discussed below). Moreover, a systematic sample is always spread out over all parts of a list. For example, if our list is ordered chronologically by the dates of transactions or records, such a sample will cover the whole time period represented in the frame.

Systematic selection is relatively simple, and commonly used. At least two potential complications can arise– the ordering of elements on the list, and dealing with fractional intervals.

Order of the List

The ordering of elements within the list can pose the most important risk in systematic sampling. The size of the fixed selection interval should not correspond with any periodicity on the list. Suppose we are studying the prevalence of different types of recreational activities, and we sample records by systematic selection from a list that sequentially orders consecutive dates. If we use an interval of 7 (or some multiple of 7), all dates in the sample would fall on the same day of the week as the first selection. Since activity patterns vary across days (Monday and Saturday activities are quite different for many), we would not want a sample of dates consisting of only one day of the week. Any interval other than a multiple of 7 would yield a good mix of days and provide a more representative picture.

Periodicity is a particularly obvious example, but other, more subtle, issues of ordering can also arise. Consider a list of persons ordered from youngest to oldest. Depending on the size of the list and the interval size, different random starts could produce samples with noticeably different age distributions. If the interval spans multiple ages, the random start will make a difference: a low random start will result in a younger sample, and a high one will produce an older sample. On the other hand, if the interval is smaller than the number of persons in the frame with any given age, the age distribution will not depend noticeably on the random start. If the highest and lowest possible random starts would fall on persons in substantively different age groups at the beginning and the end of the frame, it would probably be best to order the frame by some other variable.

If the frame cannot be reordered and the order of the list is of concern, a simple and effective approach is to change the random start as selection proceeds. With an interval of 10 and a random start of 2, for example, our first selections would be elements 2, 12, 22, 32, and so on. After reaching element 100, we could select a new random start, say 8, selecting elements 108, 118, 128, 138, and so on, until we change the random start again. This involves little more work than using a single random start.

This point anticipates a subsequent discussion of “implicit stratification.” Often a frame is deliberately sorted in a certain order to ensure that samples include all parts of a distribution. Ordering persons by age and selecting systematically ensures that we sample our “fair share” of older, middle-aged, and younger persons without creating explicit strata. Samplers like to take advantage of opportunities to stratify frames in such a simple manner. We must remain sensitive to the possible impact of the random start on a systematic sample, however, even when a list is ordered deliberately.

Fractional Intervals

Fractional intervals are the other complication in systematic sampling. If systematic selection is done by hand, it is easier to use a whole-number interval. Suppose a list contains 9,560 elements and we want to select 200, so that the interval is 9,560/200 = 47.8. A simple approach is to round fractional intervals. Rounding up lowers the sample size and rounding down raises it. The calculated interval of 47.8 in this example could be rounded up to 48, yielding 9,560/48 = 199 selections (for most random starts), or down to 47, leading to 9,560/47 = 203 or 204 selections (depending on the random start). Usually it does not matter if the sample is a little larger or smaller, especially if we have to allow for losses due to ineligibility and non-response.

If we really need to select a specific number of elements, Figure 5.2 illustrates a procedure to do so, using a fractional interval. The procedure is as follows:

- Calculate the fractional interval. To select exactly 4 elements from a list of 10, use the interval 10/4 = 2.5.
- The random start should be a fractional number greater than 0 and less than or equal to the interval. In Figure 5.2 the random start is 1.5. To obtain a fractional random start between 0.1 and 2.5, one could pick a random integer between 1 and 25 (10 times the interval), and divide by 10. For example, the random integer 15 would yield 1.5.
- Add the interval repeatedly to the random start to generate a series of selection numbers, retaining the decimal fractions, until a selection number is beyond the end of the list. In the example, the series is 1.5, 4.0, 6.5, 9.0, and 11.5.
- Truncate each selection number to a whole number by dropping its decimal portion. The truncated selection numbers in the example are 1, 4, 6, 9, and 11. Numbers that truncate to 0 and those beyond the end of the list (like the last number, 11) are discarded. Truncation is simple to do, and it yields the correct probability of selection for all elements on the list (Kish, 1965, p. 116).

(Figure 5.2 about here)

In the example, the interval between selections alternates between 2 and 3. It is 3 between 1 and 4 and between 6 and 9, but it is 2 between 4 and 6. The procedure yields exactly the desired number of selections.

Simple random sampling and systematic sampling are most commonly used to select samples in which each element in the frame has the same selection probability. Both techniques can also be applied to select elements with unequal probabilities. We next cover the most common such situation, selection with probability proportional to size.

5.3.3 Sampling with Probability Proportional to Size

Sampling with probability proportional to size (PPS) gives “larger” elements on a list a greater chance of selection than “smaller” ones. Specifically, the probability of selecting an element is directly proportional to its size. If one element is twice as large as another, it will have double the chance of being sampled.

Selecting with PPS is common in two-stage (or multi-stage) cluster samples (discussed below), in which first-stage selections are areas or other clusters that contain varying numbers of last-stage units (e.g. persons or households). First-stage units (clusters) are selected with PPS, while last-stage units are usually drawn with probability *inversely* proportional to size. PPS selection also is used for single-stage samples of units that vary in size, such as schools or businesses. In such cases, for a fixed number of selections, a PPS sample usually generates more information than a sample selected with equal probability. The PPS sample will tend to include more of the larger units than an equal probability sample in which small and large units have the same chance of selection.

Preparing the frame

In order to select a PPS sample, each element in the frame must have an associated “measure of size” (MOS). The size measure provides the basis for selecting some elements with greater probability than others. Very often the MOS is a measure of *estimated* size, so this procedure is sometimes called selection with probability proportional to *estimated* size (PPES). However, we ignore that distinction and refer to the method simply as PPS.

Figure 5.3 illustrates PPS selection. The bottom part of that figure lists 10 elements. The second column gives the measure of size associated with each element, which ranges from 1 to 7. The MOS can be in any appropriate units – population totals, sales figures, square footage, number of students, or whatever, provided that the units are the same for all elements on the frame. The scale of the units is less important than the *relative size* of the measure for different elements.

(Figure 5.3 about here)

The third column in the figure shows the cumulative running total of the MOS as we go down the list. The total of the MOSs for the 10 elements in the frame is 40 units. We calculate a selection interval using this total if we draw a PPS sample using systematic sampling.

The fourth column in the figure shows the selection range for each element—how the total range of 40 MOS units is divided among the 10 elements in the frame. The width of each element's selection range corresponds to its MOS, larger elements having wider ranges than smaller ones.

Methods of PPS selection

With selection ranges determined for the elements, we can select a sample. Because samplers usually want to minimize the chance of selecting the same element more than once, they often select PPS samples using systematic selection. However, as for an equal probability sample, we can use either simple random or systematic selection.

Simple random selection with PPS works in the same way as for equal probability samples, except that random numbers refer to the **selection range** of each element instead of its position on the list or some other identifier. The MOS of an element determines the width of its selection interval and in turn its chances of being selected. In Figure 5.3, selection ranges for all the elements together extend from 1 to 40, so the generated random numbers should lie within that range. Suppose we generate or look up the random number 5. That random number selects the element with a selection range that includes 5: element #1, with a selection range of 1 to 5. Because element #1's selection range is five times larger than element #3's (of width 1), a randomly generated number will, on average, select element #1 five times as often as element #3. Using MOSs to determine selection ranges makes the probabilities of selection proportional to the size of each element.

Systematic selection of a PPS sample works the same way as SRS selection, except that the numbers for selections are generated systematically by adding the selection interval to a random start, instead of independently.   It is important to understand that the selection interval must be based on the *total MOS*. In the example shown in Figure 5.3 we want to select three elements, so the interval is 40/3 = 13.3.   We then generate a random start between 0.1 and 13.3, say 5.5. Using the method for fractional intervals with truncation, we generate three selection numbers, 5.5, 18.8, and 32.1, which are then truncated to 5, 18, and 32, respectively. These numbers fall within the selection intervals of elements #1, #5, and #9, so those three elements are selected. Once again, letting selection intervals differ according to the MOS makes probabilities of selection proportional to size.

If an element's MOS exceeds the magnitude of the selection interval, it is certain to be selected once and might even be selected more than once. Rather than leaving such elements on a list for PPS selection, we often include them in the sample automatically as "certainty selections" and remove them from the list before sampling. In single-stage PPS samples, weights adjust for differences in selection probabilities for certainty selections. For multi-stage samples, certainty selections are treated as distinct strata, and subsamples of other units are drawn from them.

It is also possible to leave large elements on a list for PPS selection when drawing multi-stage samples, even though they must be selected at least once. This may be the most convenient approach with long lists. If a large first-stage element is selected twice, then the size of the second-stage subsample from it is doubled.

Problems can also arise if some first-stage elements are too small to yield sufficiently large second-stage samples. In such cases, groups of two or more first-stage elements can be formed. Grouped units will be selected (or not) together, with an MOS based on their combined MOSs. Kish (1965, pp. 244-245) describes a clever objective method of linking small units *after* selection, especially if they are too numerous to link by hand in advance.

We have described and illustrated the basic methods of random sampling from a single list. The next sections discuss topics involving sample *design* rather than the mechanics of drawing samples: these topics are stratification and clustering.

5.4 Stratification

Stratification is a procedure whereby we divide the sampling frame for a population into separate subpopulation frames, in order to draw a separate sample from each subpopulation. In practice, stratification usually entails dividing a big computer file up into smaller files, so that we can sample separately from each. There are several good reasons for dividing the overall frame into subpopulation frames. Unlike sample selection, however, this division is not based on some random process. We first review some reasons for stratifying, and then we show how to apply the random sampling methods of previous sections to the strata.

5.4.1 Reasons to stratify

Both theoretical and practical reasons underlie the technique of stratification. The practical considerations are usually the more decisive. The two most common reasons behind stratification are to facilitate making estimates[3] for subgroups and to increase sample precision (that is, to reduce the size of standard errors and confidence intervals).

Separate reporting areas – proportionate sampling

Research studies often seek to obtain separate estimates for parts of the population. For example, a sample of schools might need to produce results separately for different geographic regions. A reasonably large simple random sample would probably include some schools in all major regions, but it might not (because of the random selection process) contain enough schools to make adequately precise estimates for some of the smaller regions. Stratifying the frame by region and drawing separate samples would allocate a proportionate share of the total sample to each region.

[3] The term "estimate" means a particular result calculated from the sample. It is our estimate of the corresponding value in the population from which the sample was drawn.

Figure 5.4 illustrates stratification. There, a frame including 1800 schools is divided into subpopulation frames for three regions. Then a separate sample is drawn from each regional frame. Following the design in the second column, we select the *same proportion* of schools from each region, with a sampling fraction, *f*, of 0.10 or 10%. This is known as a "proportionate stratified sample."

(Figure 5.4 about here)

A proportionate stratified sample design ensures that each stratum (here, region) will be represented in the sample in proportion to its size in the population–including, in this case, exactly 10% of the schools in each region. A simple random sample from the entire frame should yield *approximately* 10% of the schools in each region, but the actual percentage in each region will vary from sample to sample. We may not want to risk ending up with a smaller than expected sample from a small stratum (like Region #2 in Figure 5.4). Stratifying guarantees that we will have a certain number of cases in each stratum. If we must report survey results separately for values of some variable, stratifying by that variable is a good idea.

Stratifying requires that information on every element's stratum be in the frame before the sample is selected. We cannot stratify on variables that will only be measured during the survey itself. Geography is often used for stratification since geographic variables are usually known ahead of time for all elements in a frame.

Oversampling some strata – disproportionate sampling

Stratifying by some variable such as region and selecting proportionately will ensure that the sample includes a certain fraction of cases from each stratum, but that may not be enough for some smaller strata. If we want good estimates for certain subgroups (strata) of the population, we may need to allocate more than a proportionate share of the sample to those strata. Having larger samples in those strata will allow us to calculate results for those strata with more precision. This type of sample is called a "disproportionate stratified sample."

The design in the third column of Figure 5.4 illustrates disproportionate stratification. The sampling fraction, *f*, differs across strata. In the figure, large Region #1 (with 1,000 schools) is sampled at a low rate (5%), small Region #2 (300 schools) is sampled at a high rate (15%), while medium-sized Region #3 (500 schools) is sampled at an intermediate rate (10%). This increases the sample size in the smaller strata, to provide enough cases to make reasonably good within-stratum estimates of the variables of interest. Limited budgets may often require reducing the sampling fraction in the bigger strata to compensate for larger samples in smaller strata.

Although disproportionate sampling improves the precision of estimates within the smaller strata, it generally reduces the precision of estimates for the overall sample, compared to a proportionate sample of the same size. Because the sample is no longer spread over all strata (regions) in proportion to the population, we need to use weights when calculating statistics describing the whole sample. These compensate for disproportionate selection, which results in having "too many" cases from smaller strata and "not enough" cases from larger strata in the sample. The consequence of having to use such weights is a reduction in precision for the overall sample.[4] Disproportionate selection involves a tradeoff between overall precision and precision in smaller strata. This tradeoff is the price we pay to have a single survey do multiple jobs. If we want reasonably good estimates for small subgroups, and if we can sacrifice some precision in the estimates for the population as a whole, then disproportionate sampling can be a good strategy.

Disproportionate sampling based on screening

Suppose we want to oversample certain ethnic groups in a population. If our frame (e.g. a list of students or hospital patients) includes a race or ethnicity code, we can create strata for the ethnic groups and sample some groups with higher sampling fractions than others. However, if we must use another frame (e.g., a list of telephone numbers or addresses) that lacks ethnicity data, we cannot stratify ahead of time. Instead we must begin the interview with "screening" questions, to ascertain the ethnicity of those selected, and then oversample by continuing with the full interview at different rates for

[4] See Kish, 1965, pp.429-431, for a method to estimate the loss in precision due to the oversampling of strata. Software to calculate complex standard errors for specific variables will automatically include the effect of weighting in the standard errors, but Kish's procedure offers a convenient way to isolate and estimate the overall effect of weighting for a particular survey design.

different groups. For instance, we might interview all African Americans and Latinos in a sample, but only half of those in other groups.

Fieldwork planning and supervision must control the implementation of screening procedures like this "continue half of the time" rule. Our preference is to control such selection rules by dividing the sample into random parts ("replicates") and then assigning a different selection rule to each part. For the example in the preceding paragraph, we would divide the sample at random into two halves. In one half, interviewers would attempt to complete the interview with everyone. In the other half, they would attempt to interview only African Americans and Latinos. African Americans and Latinos would then have double the probability of selection into the overall sample, compared with the other groups.

Reducing sampling error – "optimal allocation"

Often a major reason for stratifying is to attempt to increase the precision of statistics by creating strata based on one or more variables that are correlated with the primary variable we are trying to estimate. If the variation of our primary variable within strata is less than its variation overall, *proportionate* stratification will increase the precision of the estimate of our primary variable (see Groves et al., 2009: pp. 114-120; Kalton, 1983: pp. 20-24).

Disproportionate stratification can sometimes be used to increase precision even more, by using a strategy called "optimal allocation" (see the Frankel and the Land and Zheng chapters in this volume). Optimal allocation is a strategy for allocating more (than proportionate) cases to those strata with relatively high variability in the primary variable of interest. Specifically, if data collection costs are the same in all strata, the sampling fractions in the strata should be proportional to the primary variable's standard deviation in each stratum. For instance, if the primary variable's standard deviation is twice as large in stratum #1 as in stratum #2, the sampling fraction in stratum #1 should be double the sampling fraction in stratum #2.

If data collection costs differ across strata, optimal allocation also calls for increasing the sampling fraction in low-cost strata, and decreasing it in more expensive strata. More specifically, sampling fractions should be inversely proportional to the square root of the cost per case in a stratum. For example, if costs per case are four times greater in one stratum compared to a second, the more expensive stratum should be sampled at half the rate.

The combined criteria of variability and cost can be summarized as:

$$f_h = k * S_h / \sqrt{C_h}$$

where f_h is the sampling fraction in stratum h, S_h is the standard deviation in stratum h of the primary variable to be estimated, C_h is cost per element in that stratum, and k is a constant used to scale the sampling fractions to produce the target sample size.

When these criteria can be applied, sampling theory shows that confidence intervals for means, percentages, and totals based on the overall sample will be as small as possible for a given budget (Kish 1965, pp. 92-98; Kalton 1983, pp. 24-26).

Unfortunately we often lack the information necessary for applying those optimization criteria. Unless estimates are available from prior studies, we may not know the details of the primary variable's distribution in advance, and will not be able to estimate its standard deviation in various strata. Moreover, costs per case are often little different for different parts of the frame.

And finally, one rarely conducts a whole survey just to obtain estimates for a single variable. Surveys are almost always multi-purpose, and the optimal sample allocation for one variable may not be optimal for some other variable of equal importance. Proportionate stratified sampling, with the same sampling fraction for all strata, is usually best – unless we have a good reason to oversample a particular subgroup.

Nevertheless, optimal allocation is a very helpful heuristic for designing a sample. Stratification is not simply a matter of convenience or a way of producing reports for separate parts of the sample. The goal of good sample design is to generate samples that produce results that are as precise as possible, and stratification helps to do that. It is among the most useful tools available for designing samples.

5.4.2 Methods of stratification

Stratification may be achieved *explicitly* by creating sub-frames, or *implicitly* by exploiting the order of elements in a single frame. Some sample designs combine the two.

Explicit stratification

In introducing stratification, we tacitly assumed that strata are created explicitly, by physically dividing the overall frame into separate sub-frames or files. Then a separate sample is drawn from each. This is the basic method of stratification. No formulas dictate how many strata to create. From a practical point of view, the number of strata required depends on the number of separate subgroups for which results must be presented and on whether we can subdivide the population based on a variable that is correlated with the variable(s) of primary interest.

If we plan to use disproportionate stratified sampling, we must keep track of the relative sampling fractions for strata, so that the strata can be weighted appropriately to reflect the population. Then we will be able to use those weights to combine the data from different strata when calculating results for the overall sample, If, on the other hand, we do not plan to apply different sampling fractions to different parts of the frame, we do not always need to stratify explicitly. A simpler method, implicit stratification, is often sufficient.

Implicit stratification

Stratifying a frame before sample selection ensures that the sample is distributed over the various segments of the population. “Implicit stratification” accomplishes this without creating explicit strata for the various segments.

With implicit stratification, we sort the frame by some variable and then select a systematic random sample. For example, to ensure that a sample of addresses is spread over all regions of a state, we could first sort the address list by zip code, and then select addresses with systematic sampling (**not** with SRS, which would defeat the purpose of sorting). By selecting the sample in this manner, we can be sure that the sample will include addresses from all of the major geographic areas included in the frame.

Spreading the sample over the distribution of a variable may also improve the precision of the statistics we are estimating. In a study of health variables, for instance, sorting a frame of persons by their age will usually be helpful, since age is highly correlated with health status. Controlling the age distribution in the sample should therefore reduce the sampling error of estimated health statistics.

Stratifying implicitly is often more practical than stratifying explicitly. Creating explicit strata for zip code groups, for example, could require a fair amount of work: examining the distribution of elements in the frame by different series of zip codes, deciding how many strata to create, and finally dividing the frame into separate files. Sorting by zip code is much easier than going through all those steps.

Another reason to stratify implicitly on a variable is that we might prefer to base explicit strata on other variables. For example, we might need to stratify a list of schools by type of school (public, private, charter) and by grade level. Creating explicit strata for groups of zip codes would reduce our opportunity to stratify on these other important variables. It may be preferable to sort on zip code *within* explicit strata defined by the other variables. We comment further below on this very useful combination of explicit and implicit stratification.

Implicit stratification is very useful and common, but it cannot achieve all the goals of stratification. In particular, using disproportionate stratification to oversample certain subgroups requires the creation of explicit strata so that a larger sampling fraction can be applied in certain strata. Also, implicit stratification cannot guarantee a specific number of selections in any particular segment of the frame. Explicit strata should be created if this is important for reporting results. Finally we should check for ordering effects in any systematic sample. If the selection interval is large compared to the number of elements in each category of the variable we are sorting on, high or low random starts could produce samples that differ in non-random ways.

Combining explicit and implicit stratification

Stratification imposes some control on the sample selection process by ensuring that a sample is spread over the distributions of certain variables in a predictable way. In general, more strata yield better control. Consequently, samplers tend to stratify the sampling frame as much as they can.

It is often desirable to stratify by more than one variable at the same time (for instance, by creating a stratum for each school type within each region). Explicit stratification offers the most control over sample selection, but a frame can be divided into only so many categories at once. A solution is to create explicit strata based on some variables, and then sort the frame on other variables *within* each explicit stratum, to gain the benefit of some additional implicit stratification. This combination of explicit and implicit stratification is common.

Explicit stratification is often used for major geographic areas such as regions or states, especially if we know in advance that separate results will be required for those segments of the population. If information for further stratification is available in the frame, the simple device of sorting on one or more variables and then selecting systematically within each explicit stratum takes advantage of additional opportunities to attain the goals of stratification.

5.5 Cluster Sampling

When we sample, our eventual goal is to collect data on a specific type of "element" (e.g., students). An "element sample" selects elements directly, as from a list. So far, everything in this chapter has been about "element sampling." Often, however, we plan to sample elements only though *groups of elements* known as "clusters," usually to reduce costs. Such circumstances require "cluster sampling."

Figure 5.5 presents an example of a cluster design for sampling students in a state. Often we cannot sample students (the elements) directly, because listing them would be too costly, or because we wish to concentrate the sample in a limited number of schools to reduce costs during data collection. So instead of selecting students directly, we might select students within a sample of schools (clusters). Within each selected school we will select some (or all) of the students. In the figure, School #1 and School #3 are selected as clusters for further sampling of students, but School #2 and School #4 are not.

(Figure 5.5 about here)

Because the same groups of elements (like schools) could be used either as strata or as clusters, the distinction between stratification and clustering can be confusing. Strata and clusters differ in an important way. After dividing the elements in a frame into *strata*, we subsequently sample elements *from every stratum*. The point of grouping elements into *clusters*, however, is that we select elements *only from some of the clusters.*

Effect of cluster sampling on precision

Cluster sampling usually increases the size of standard errors and confidence intervals of the statistics we calculate from the sample results. Notice in Figure 5.5 that we will not sample any students in schools #2 and #4. Nevertheless, we certainly will want to generalize results to all students in the state– not only to students in those schools that happen to have been selected as clusters. Since clusters are selected at random, the results can be generalized to the whole population, but the sampling of clusters introduces a new level of uncertainty into our results.

What if we had selected, by chance, other clusters into the sample – how different would the study results be? How different are the clusters (schools) of students from one another, in regard to the variables we want to study? If the sampled schools are not very different, we can reasonably infer that our results would have been similar had we sampled other schools instead. If, on the other hand, the sampled schools turn out to be quite different from one another, our uncertainty due to the sampling of clusters increases, which correspondingly increases the width of confidence intervals for statistics based on the sample. Campbell and Berbaum (this volume) cover methods of computing these confidence intervals for cluster samples; here we try to provide an intuitive understanding of the issues.

Comparing two extreme cases is informative. Consider a sample of 2,000 students within 100 schools, an average of 20 students in each. Suppose that some characteristic (a certain test result, for instance) of *all students within each school is exactly the same,* but the results for all sampled schools differ from one another. In this case, all the information about test results in a school could have been obtained from a single student in each school. Instead of sampling 2,000 different students, we could have learned just as much from only 100 students, with one student per school. So our cluster sample of 2,000 students is the equivalent of a simple random sample of only 100 students. Calculating a confidence interval by assuming that we have a simple random sample of 2,000 independent selections overstates sample precision, because of the high (here, perfect) correlation between elements within clusters. When elements within clusters are homogeneous, sampling additional elements within clusters provides less information than one might expect.

Now consider the other extreme case. Consider the same sample of 2,000 students within 100 schools. What if the *average* of some characteristic (e.g., a certain test result) was *exactly the same for all schools*, though students *within* schools differed from one another on that characteristic? Then there would be no "cluster effect" on the results; it would have made no difference if we had sampled 2,000 students from 100 schools, or 40 schools, or even 2 schools (if they were large enough). In this ideal case, the cluster sample of 20 students within each of 100 schools is equivalent to a simple random sample of 2,000 students from a statewide list. Both samples would have the same confidence intervals. This is ideal: we conserve resources by dealing with only 100 schools, but we obtain results as precise as those from a sample of 2,000 students spread around the state.

In reality, of course, the effect of clustering almost always lies somewhere between these two extremes. Results usually differ between clusters, and rarely are all elements within clusters exactly the same. The more the variability *between* clusters and the less variability among elements *within* clusters, the lower the precision of sample statistics in a cluster sample.

Understanding the tradeoffs

Cluster sampling involves a tradeoff between sample precision and data collection cost. From a precision standpoint, no clustering at all is best, and spreading the sample over many clusters is preferable to concentrating it within a few. Statistics will almost always be more precise for a sample of 500 elements with 100 clusters and 5 elements in each cluster than for 25 clusters with 20 elements in each.

Usually, however, the design with more clusters will cost substantially more. Gathering data in a new cluster may involve additional travel expense and other costs (e.g., additional time negotiating access with a new school principal or other "gatekeeper"). Such costs are greater than those of collecting data on an additional element within an already-selected cluster. If the cost of including an additional cluster (school) is 10 times that of collecting data on an additional element (student) within an existing cluster, the "relative cost" is 10.

By using fewer clusters and increasing the number of elements in each one, we can afford to collect data on more elements, which is one way to increase precision. At the same time, concentrating the sample elements in fewer and bigger clusters will usually reduce precision. How can we balance these conflicting goals – high precision at minimal cost – into a workable design?

Cluster effect and design effect

Quantifying the "cluster effect" can help us resolve this tradeoff.[5] Sampling theory calls this effect the "coefficient of intraclass correlation" and represents it by *roh* or the Greek letter ρ. Kish (1965, p. 161) clarifies by calling it a "rate of homogeneity." Like the familiar Pearson correlation coefficient, *roh* is scaled to range between zero and one.

We can calculate *roh* only after a study is completed and standard errors have been computed (as discussed in by Campbell and Berbaum, this volume). When designing a cluster sample, however, it is useful to have a guess about the probable size of *roh*, perhaps based on results of other studies that used similar samples. Most research reports do not present values of *roh* itself, but they sometimes report the "design effect," from which we can calculate *roh*.

The design effect, *deff*, is the ratio of the variance of a statistic calculated from a cluster sample (or any complex sample) to that of the same statistic calculated from a simple random sample of the same size. For example, if the variance of a statistic in a cluster sample is twice as large as its variance under SRS, the design effect is 2.

The following important formula (Kish 1965, pp.161-164; Groves et al. 2009, pp. 109-112) gives the relationship between *roh* and *deff*, where *b* is the average number

[5] See Frankel (this volume) for more detail on cluster and design effects. Harter et al. (this volume) also discuss these issues.

of elements per cluster:

$deff = 1 + roh(b - 1)$

As the formula makes clear, we can reduce the design effect, and improve precision, either by using clusters that have a low *roh* (low homogeneity), or by designing a cluster sample with a low cluster size *b*. If *roh* is zero, the design effect will be 1 regardless of the cluster size *b*. But if *roh* is high, even a relatively small cluster size will result in a high *deff*.

Solving for *roh* in terms of *deff* and *b* yields:

$roh = (deff - 1) / (b - 1)$

If a study similar to ours reports design effects and provides the information needed to calculate average cluster size (the total number of elements and the number of clusters), we can calculate *roh* and use that information to design our cluster sample.[6] Or, if we have access to the data file of a prior study, we can calculate *deff* and *roh* for ourselves, using newer versions of statistical packages like Stata or SAS that calculate the correct variances and standard errors for cluster samples.

In any case, to optimize the design of a cluster sample we must make some guess about the value of *roh* that we expect to encounter. In some studies *roh* is relatively small, like 0.05. A moderate *roh* is 0.10, and a high one is 0.20. Notice that even a moderate *roh* of 0.10 will produce a *deff* of 2 if the average cluster size is 11, so that the confidence intervals for the cluster sample will be 40% wider than those for a simple random sample of the same size. (If the variance is two times larger, standard errors are larger by the factor √2 = 1.4)

Optimal cluster size

With an estimate of *roh* for the primary variable of interest in a sample that uses a specific type of cluster design, we can begin to resolve the precision-cost tradeoff described above. We also require information on the relative cost of adding a new cluster versus collecting data from one more case in an already selected cluster. An easy-to-

[6] Design effects may be reported in two different forms. One form is *deff*, the ratio of the *variances* of a statistic from a cluster sample and from a simple random sample of the same size. The other form, *deft,* is the ratio of the *standard errors* (the square roots of the variances) of the statistic for the two types of

apply formula gives the optimal cluster size, *b,* for a given *roh* and relative cost (Kish, 1965, p. 269):

optimal $b = \sqrt{\text{relative cost} * (1-roh)/roh}$

For example, with a *roh* of 0.05 and a relative cost of 10, the optimal *b* is √(10*19)=14 (rounded). This means that we should plan to sample about 14 elements per cluster. That degree of clustering should produce the narrowest confidence intervals possible for a given budget, for those variables having a *roh* of 0.05. Precision will be lower for variables with a higher *roh,* and greater for those with a lower *roh*. Table 5.1 gives the optimal cluster size for several combinations of relative cost and *roh*. Notice that only when relative cost is very high or *roh* is very low do larger cluster sizes give the optimal result.

(Table 5.1 about here)

Different variables can and do have different values of *roh,* and therefore different optimal cluster sizes. Moreover, we are often guessing about the size of *roh*. In practice, then, the cluster size is often set using a compromise figure. Nevertheless, the exercise of calculating optimum cluster size has heuristic value for designing good samples, by requiring us to think systematically about the tradeoffs. Reducing costs is not the sole object of cluster sampling. For any given budget, we want a sample design that provides the most precise results possible.

Selecting clusters

Selecting clusters requires a frame of clusters, and uses the techniques already described above for selecting individual elements from a frame. As a first step, it can be advantageous to stratify clusters, to ensure that the selected clusters are spread over the whole population. We may also plan to oversample certain strata (types of clusters). Stratification of clusters could also reduce sampling error, if the clusters can be grouped into strata likely to differ on the variables of interest, since the standard errors for statistics will be computed based on differences between clusters *within the same*

samples. Since *deft* is the square root of *deff,* if *deft* is reported one should convert it to *deff* by squaring before using the formula in the text to calculate *roh.*

stratum. Through such stratification, we might mitigate some of the loss of precision that usually results from cluster sampling.

Cluster sampling can be carried out either as a one-stage sample or as part of a two-stage (or multi-stage) sample. An example of a one-stage cluster sample is a sample of students within schools, in which we collect data on **all** students within the selected schools. One-stage samples have large clusters, and usually large design effects as well, so confidence intervals for most statistics will be wider than one might expect for the number of students sampled.

Nevertheless, the type of data involved, and the cost structure for collecting them, may justify sampling complete clusters. Suppose that the main cost of a survey of students is the initial cost of contacting a school and getting access to its records. After that, the marginal cost of data on additional students within that school may be negligible, especially if the data are computerized. That is, the relative cost of selecting an extra cluster (school), compared to that of collecting data on an individual element (student), may be so high that it justifies large clusters even with a high expected *roh*.

Two-stage cluster sampling

Often, however, we want to sample only some of the elements in the selected clusters. Then we need a two-stage sample. A certain number of clusters are selected in the first stage, and then elements are selected only within the selected clusters in the second stage. Clusters are stepping stones providing access to the elements within each cluster. Large-scale area probability samples (Harter et al., this volume) are an important application of such designs. We briefly discuss their use in smaller scale studies here.

In two-stage cluster sampling, one should decide on the selection method for the two stages jointly. The simplest method is to select clusters with **equal probability** at the first stage, and then to select elements, also with equal probability, within the selected clusters. This method produces an equal-probability sample that would not require sampling weights to be used in analyses. For example, we could select 1% of the schools in a state and then subselect 10% of the students in each selected school. The overall probability of selection would be 1/100 x 1/10 = 1/1000 and would be the same for all students in the state. However, this method yields little control over the total sample size. If the selected clusters happen to be larger schools, the 10% subsamples will also be large; if they happen to be small, the 10% subsamples will be correspondingly small. Stratifying the schools by size could control the sample size to some extent, but then we give up the opportunity to stratify using some other, perhaps more interesting, variable(s).

A more efficient way of maintaining control over sample size is to sample clusters with **probability proportional to size** (PPS), and then to subsample elements within the selected clusters with probability ***inversely*** **proportional to size**. Suppose we plan to select 5 elements per cluster. If the first-stage PPS sample selects a cluster with a measure of size (MOS) of 100, we would subsample elements within it with the fraction 5/100: either sampling elements at random at the rate of 5%, or systematically sampling them using an interval of 20 and a random start between 1 and 20. Element samples within each of the other selected clusters would be drawn using a fraction based on its respective MOS – that is, 5 / MOS_i. This procedure can be summarized with the following equation:

$$\text{Probability} = (a * \text{MOS}_i \ / \ \text{Total\_MOS}) \ * \ (5 / \text{MOS}_i)$$

where MOS_i is the measure of size for cluster *i*, and *a* is the number of clusters selected.

Sampling with PPS at the first stage and inverse PPS at the second stage produces an equal-probability sample. Notice that the MOS_i in the equation above then cancels out: the overall sampling fraction (or probability of selection) is the same (i.e., 5a/Total_MOS) for all elements in all clusters. Therefore it is not necessary to use sampling weights in analyses. The advantage of this method is that total sample size is quite predictable, provided that the actual cluster sizes found later during fieldwork are not very different from the MOSs for the clusters. To ensure that the overall sample remains equal-probability, subsampling from each selected cluster must be based on its MOS, *not* its actual number of elements found later during fieldwork (otherwise the MOS_i in the equation above will not cancel out).

If we decide to select exactly 5 units in a cluster (instead of applying the second-stage fraction 5/MOS_i), our second-stage sampling fraction will be 5/N_i where N_i is the actual number of units in the cluster found during fieldwork. Then the overall probability of selection would be:

Probability = $(a * MOS_i \ / \ Total\_MOS) \ * \ (5 / N_i)$.

Notice that MOS_i and N_i do not cancel each other out of this equation, unless they are exactly the same in every cluster (which is unlikely). The units selected in cluster i would therefore be selected with probability proportional to the ratio MOS_i / N_i which could be different for every cluster. We should compensate for such a departure from equal-probability sampling by using weights, a topic we turn to next.

5.6 Weighting

Several features of samples, even for small-scale studies, may require that weights be used in data analysis. This section provides a brief summary of the principles of weighting.

Weights give some cases more influence (weight) than others when calculating statistics. Their basic purpose is to correct for biases in the data, resulting from either the sample design or data collection procedures, that end up producing "too many" sample elements from one population segment, and "not enough" from some other segments. The sample designer should provide instructions for creating basic sampling weights for any sample design other than an equal-probability sample.

Relative weights versus expansion weights

One distinction cuts across all types of weights: that between relative weights and expansion weights. This difference is simply a matter of scale.

Expansion weights scale the total weighted number of cases up to the size of the population that the sample represents. For example, if we sampled 1% of students from some list, each student would be given a weight of 100 (on average). If that 1% sample yielded 500 students, the expansion weights would project sample results up to the 50,000 students in the population. Expansion weights are especially useful when presenting results to policymakers or other publics interested in knowing not only *what percentage* of people have some characteristic but also *how many*.

Relative weights scale the weighted number of cases to the actual size of the sample, and they usually have a mean of 1. Some cases have relative weights greater than 1, and others have relative weights less than 1, but the total weighted number of cases is the same as the actual sample size. Data analyses and presentations of results often use relative weights, to convey an approximate sense of the precision of sample statistics. Using expansion weights could give the misleading impression that statistics are based on tens of thousands of cases, when in fact the sample may only include a few hundred.

Expansion and relative weights for different cases in a given sample should have the same proportionality to one another. For example, one case might have a relative weight of 1.5, and another a relative weight of 0.75. The corresponding expansion weights might be 1,000 and 500 – in the same ratio of 2:1. When calculating descriptive statistics other than totals, using either type of weight should give the same results. All weighting adjustments discussed below can be used to construct both expansion weights and relative weights. Expansion weights can readily be converted into relative weights by dividing them by the mean of the expansion weights. To convert a relative weight into an expansion weight, we must know the total population size or the sampling fraction.

Adjusting for selection probabilities

Section 5.4 introduced disproportionate stratified sampling, in which we divide a sampling frame into several strata and sample the strata at different rates. For instance, with a sampling frame divided into geographic regions, we might sample smaller regions at higher rates than larger ones, to increase the sample size and thus the precision of estimates in smaller regions.

It is crucial to keep track of the sampling rate used in each stratum. When we combine results from different strata into estimates for the full population, data from different strata must receive different weights to take into account the oversampling of some strata and the undersampling of others. This first weighting adjustment factor, applied to every case in the data file, is based on the inverse of the sampling fraction in each case's stratum:

Weight factor #1 = $1/f_h$

where f_h is the sampling fraction for stratum *h*. If we sample elements in stratum 1 with the fraction 1/100, and those in stratum 2 with the fraction 5/100, the first weight factor for the cases in stratum 1 will be 100, and the factor for stratum 2 will be 100/5 = 20.

Sometimes the information needed to adjust for different probabilities of selection is only available after the fieldwork has been completed. For example, in household samples of adults, usually only one adult is selected at random to be interviewed within each sampled household. An adult who lives alone will always be selected if we select her or his household. In comparison, the chance of selecting an adult who lives with one other adult is only half as large. However, we do not know the number of adults in the household until after it is selected and contacted.

Differences in selection probabilities for households due to multiple telephone numbers in random-digit-dialed telephone samples are another common example. A household with two separate telephone numbers (regularly answered and not used exclusively for a fax machine or a computer modem) has twice the chance of selection as one with a single telephone number. Likewise, if cell phone numbers as well as landline numbers are in the sampling frame, they also affect the probability of selecting individuals. Someone who receives calls via a cell phone has one chance to be called on the cell phone, and another to be selected through the household's landline. Whenever the elements in the survey population are selected at different rates, we must compensate by using another weighting factor. This adjustment requires that the survey obtain data on the source of differences in selection probabilities (e.g. the number of adults in a household, and the number of telephone numbers). This second weighting adjustment factor is

Weight factor #2 = 1/$\mathrm{p_i}$,

where p_i reflects influences on selection probabilities for case *i*.

This weight factor can combine more than one factor affecting differential selection probabilities. If, for example, a household has two telephone lines and three eligible adults, the value of the combined value of p_i for an adult in that household is 2/3, the product of the telephone factor of 2 and the adults factor of 1/3. Since weight factor #2 is the inverse of p_i, the second weighting adjustment for such an adult would be 1/(2/3) = 3/2 = 1.5.

Non-response adjustments

Survey response rates are rarely 100%. Not adjusting for differential non-response tacitly assumes that all non-respondents are similar to the average respondent with respect to the variables measured. If non-response is concentrated in certain subgroups, statistics for the sample will under-represent those groups. Weighting adjustments for non-response compensate for this. Such adjustments assume that non-respondents in a subgroup are more like the respondents in that subgroup than the average respondent. If the subgroup classification is related to the variables we are estimating, a non-response adjustment may improve our estimates.

To make a weighting adjustment for non-response, we must calculate a separate response rate for each subgroup. In order to do that, we must know the subgroup membership for all elements in the sample – non-respondents as well as respondents. We cannot use a subgroup classification to adjust for non-response if it becomes known only after fieldwork. For example, we usually do not know the race or ethnicity of sampled persons before interviewing them, so we cannot usually calculate separate response rates for race/ethnicity subgroups. Sampling strata, therefore, are commonly used subgroup classifications for purposes of non-response adjustment, since we know the stratum membership for every sampled element.

Weighting adjustment factors for non-response are the inverse of a subgroup's response rate:

Non-response factor = $1/\mathrm{rr}_\mathrm{g}$

where rr_g is the response rate for group g, expressed as a proportion, like 0.50 or 0.45.

If response rates are similar in all subgroups, this non-response adjustment factor will also be similar for all subgroups, and it will have little or no impact on the relative size of weights. It will, however, increase the weighted number of cases. That can be important when creating expansion weights, to estimate the *numbe**r*** of elements in the population having a certain characteristic.

Putting the factors together

After calculating the factors that adjust for differences in probabilities of selection and non-response, a weight variable is constructed by multiplying them together. The value of the weight variable for case *i* in stratum *h* and subgroup *g* in the sample is the product of the factors described above:

$$\text{weight}_{ghi} = (1/f_h) * (1/\mathrm{p_i}) * (1/rr_g)$$

where:

f_h is the sampling fraction for elements in stratum *h*, and

p_i is the probability factor for selecting element *i*, as learned during fieldwork, and

rr_g is the response rate for elements in group *g*.

This weight will be an expansion weight if the sampling fractions have been expressed in absolute terms (like 1 / 10,000) instead of relative terms (for example, that stratum 1 was sampled at double the rate of stratum 2). Relative weights that yield the same number of weighted cases as the actual number of completed cases in the data file (*n*) can be calculated by dividing the above-calculated $weight_{ghi}$ for each case by the mean of the weights:

$$\text{relative weight}_{ghi} = \text{weight}_{ghi}/(\Sigma(\text{weight}_{ghi})/n)$$

This weight (either expansion or relative), adjusting for selection probabilities and response rates, is sufficient for many studies. Sometimes, however, we want to go further and adjust the sample distributions to match some criterion distribution. We turn to that topic next.

Post-stratification weights

After the weighting adjustments for selection probabilities and response rates have been made, noticeable differences between the distributions of certain variables in the sample and in the population may still exist. One common difference is for the percentage of women in the sample to exceed that in the population. The response rate is generally a little higher among women than among men, but we usually cannot adjust for differential non-response by gender because the gender of respondents becomes known only during fieldwork.

Another reason that a sample distribution may differ from a criterion distribution like the U.S. Census is that the sampling frame may not cover some groups as well as others. Race and ethnic distributions could diverge from Census figures because the sampling frame is less apt to include very low income households (because they are less likely to have telephones, for instance), and those missing households might be concentrated in particular ethnic groups.

Post-stratification weighting adjustments make the distributions of key variables in the sample match Census figures or some other criterion distribution. Matching the distributions of several different variables at once (e.g. gender, age, education, race, and income) can be quite complicated.[7] But post-stratification on one or two variables, each with only a few categories, is not difficult. Simply follow these steps:

A. Calculate the percentage of cases *in the sample* within the categories you want to adjust. For example, we could use the percentage of respondents in each cell of the cross-tabulation of race by gender. The percentages must add up to 100%. Be sure to use the weight for differential selection probabilities and non-response when generating those percentages[8], and use at least a few decimal places. Also, you should have at least about 20 cases in each cell; otherwise, use fewer categories.

B. Find the corresponding percentages *of the population* in those same categories, from Census data or some other criterion source. These too must add up to 100%.

C. For each category in the distribution, divide its population percentage (B) by its sample percentage (A). This ratio is the post-stratification adjustment factor that applies to all cases in that category. For example, making the gender distribution for the sample match the Census distribution could require adjustment factors like 1.1234 for males and 0.8902 for females. This would

[7] See Frankel (this volume) on adjusting for several variables by a process called "iterative marginal weighting," often referred to as "raking."

[8] If it is not necessary to weight for differential probabilities of selection and/or non-response, then such weights are effectively 1.0 for each case, and the unweighted percentages can be used for this step.

have the effect of increasing the weighted number of males in the sample, and decreasing the weighted number of females.

D. Finally, produce a new weight for each case, *i*, by multiplying the previous weight variable by the post-stratification adjustment appropriate to that case:

$$\text{post-stratification weight}_{ghi} = \text{post-stratification adjustment}_{i} * \text{weight}_{ghi}$$

Since the post-stratification weight includes all the adjustments incorporated into the previous weight variable, it would usually be used as the primary weight variable when analyzing the data.

REFERENCES

Groves, R. M., F. J. Fowler, Jr., M. P. Couper, J. M. Lepkowski, E. Singer, R. Tourangeau (2009), *Survey Methodology*, 2nd ed., John Wiley, New York, NY.
[Excellent basic textbook on survey research methods. It includes a good introductory chapter on sampling.]

Kalton, G. (1983), *Introduction to Survey Sampling*, Sage Publications, Thousand Oaks, CA.
[Short but good introduction to survey sampling. It includes some basic equations that are useful to have for reference.]

Kish, L. (1965), *Survey Sampling*, John Wiley, New York, NY. (newly released in 1995).
[Technical but practical guide to every aspect of survey sampling. Samplers often regard this as their “bible” because it is such a useful resource.]

Figure 5.1

Simple Random Sampling From a List

Want to select 2 out of 10 elements

Generate a few random numbers between 1 and 10:

8
4
7
6
6

List of elements	**Selected?**
Element 1	
Element 2	
Element 3	
Element 4	**Yes**
Element 5	
Element 6	
Element 7	
Element 8	**Yes**
Element 9	
Element 10	

Formula (in Excel) for generating a random number between 1 and 10:

```
=INT(RAND()*(10-1) + 1)
```

Figure 5.2

Systematic Random Sampling with a Fractional Selection Interval

Number on the list:	**10**
Number to select:	**4**
Selection interval	**2.5**
Random start:	**1.5**

Selection series:	**With fractions**	**Truncated**
	1.5	1
	4.0	4
	6.5	6
	9.0	9
(beyond end of list:)	11.5	11

Figure 5.3

Systematic Selection with Probability Proportional to Size

Total of size measures:	**40**
Number to select:	**3**
Selection interval:	**13.3**
Random start:	**5.5**

Selection series:	**5.5, 18.8, 32.1**
Truncated:	**5, 18, 32**
Method:	**Fractional interval with truncation**

Elements	**Measure of Size**	**Cumulative MOS**	**Selection Range**	**Selected?**
1	**5**	**5**	**1- 5**	**5**
2	2	7	6-7	
3	1	8	8	
4	3	11	9-11	
5	**7**	**18**	**12-18**	**18**
6	6	24	19-24	
7	2	26	25-26	
8	5	31	27-31	
9	**6**	**37**	**32-37**	**32**
10	3	40	38-40	

Figure 5.4

Stratification

	Proportionate sampling	Disproportionate Sampling
STRATIFIED FRAME		
Region 1 (large)		
School 1		
School 2		
School 3	f = 10%	f = 5%
…		
School 1000		
Region 2 (small)		
School 1		
School 2		
School 3	f = 10%	f = 15%
…		
School 300		
Region 3 (medium)		
School 1		
School 2		
School 3	f = 10%	f = 10%
…		
School 500		

Figure 5.5

Cluster Sampling

ELEMENTS WITHIN CLUSTERS	Selected?
School 1	**Yes**
Student 1	
Student 2	
Student 3	
…	
Student 190	
School 2	**No**
Student 1	
Student 2	
Student 3	
…	
Student 215	
School 3	**Yes**
Student 1	
Student 2	
Student 3	
…	
Student 350	
School 4	**No**
Student 1	
Student 2	
Student 3	
…	
Student 220	

Table 5.1

Optimum Cluster Size

			Roh			
Relative Cost	0.01	0.02	**0.05**	0.10	0.15	0.20
1	10	7	4	3	2	2
2	14	10	6	4	3	3
3	17	12	8	5	4	3
4	20	14	9	6	5	4
5	22	16	10	7	5	4
6	24	17	11	7	6	5
7	26	19	12	8	6	5
8	28	20	12	8	7	6
9	30	21	13	9	7	6
10	31	22	**14**	9	8	6
11	33	23	14	10	8	7
12	34	24	15	10	8	7
13	36	25	16	11	9	7
14	37	26	16	11	9	7
15	39	27	17	12	9	8
20	44	31	19	13	11	9
50	70	49	31	21	17	14
100	99	70	44	30	24	20
500	222	157	97	67	53	45
1000	315	221	138	95	75	63
1500	385	271	169	116	92	77

For example: If *roh* is .05 and the relative cost is 10, the optimal cluster size is 14.

Simple cost model: Total Cost = a * (cost per cluster) + n * (cost per case)
where *a* = number of clusters, and *n* = number of interviews or cases

Relative cost = (cost per cluster) / (cost per case)

Optimal cluster size = sqrt((relative cost) * **(1 - *roh*)/*roh*)**

See Kish, 1965, equation 8.3.7, p. 269.