Sampling from Finite Populations

Jill M. Montaquila and Graham Kalton
Westat
1600 Research Blvd., Rockville, MD 20850, U.S.A.

Keywords: Survey sampling, finite populations, simple random sampling, systematic sampling, stratified sampling, multistage sampling, two-phase sampling, multiple frame sampling

1. Introduction

The statistical objective in survey research and in a number of other applications is generally to estimate the parameters of a finite population rather than to estimate the parameters of a statistical model. As an example, the finite population for a survey conducted to estimate the unemployment rate might be all adults aged 18 or older living in a country at a given date. If valid estimates of the parameters of a finite population are to be produced, the finite population needs to be defined very precisely and the sampling method needs to be carefully designed and implemented. This entry focuses on the estimation of such finite population parameters using what is known as the *randomization* or *design-based approach*. Another approach that is particularly relevant when survey data are used for analytical purposes, such as for regression analysis, is known as the *superpopulation approach* (see the entry *Superpopulation Models in Survey Sampling**).

This entry considers only methods for drawing probability samples from a finite population; *Nonprobability Sampling Methods** are reviewed in another entry. The basic theory and methods of probability sampling from finite populations were largely developed during the first half of the twentieth century, motivated by the desire to use samples rather than censuses to characterize human, business, and agricultural populations. The paper by Neyman (1934) is widely recognized as a seminal contribution because it spells out the merits of *probability sampling* relative to purposive selection. A number of full-length texts on survey sampling theory and methods were published in the 1950's and 1960's including the first editions of Cochran (1977), Deming (1960), Hansen, Hurwitz, and Madow (1953), Kish (1965), Murthy (1967), Raj (1968), Sukhatme et al. (1984), and Yates (1981). Several of these are still widely used as textbooks and references. Recent texts on survey sampling theory and methods include Fuller (2009), Lohr (2010), Pfeffermann and Rao (2009), Särndal, Swensson, and Wretman (1992), Thompson (1997), and Valliant, Dorfman, and Royall (2000).

Let the size of a finite population be denoted by N and let Y_i $(i= 1, 2, ...,N)$ denote the individual values of a variable of interest for the study. To carry forward the example given above, in a survey to estimate the unemployment rate, Y_i might be the labor force status of person (element) i. Consider the estimation of the population total $Y=\Sigma_i^N Y_i$ based on a probability sample of n elements drawn from the population by sampling without replacement so that elements cannot be selected more than once. Let π_i denote the probability that

element i is selected for the sample, with $\pi_i > 0$ for all i, and let π_{ij} denote the probability that elements i and j are jointly included in the sample. The sample estimator of Y can be represented as $\hat{Y} = \Sigma_i^N w_i Y_i$ where w_i is a random variable reflecting the sample selection, with $w_i = 0$ for elements that were not selected. The condition for $\hat{Y}$ to be an unbiased estimator of Y is that $E(w_i) = 1$. Now $E(w_i) = \pi_i w_i + (1-\pi_i)0$ so that for $\hat{Y}$ to be unbiased $w_i = \pi_i^{-1}$. The reciprocal of the selection probability, $w_i = \pi_i^{-1}$, is referred to as the *base weight*. The unbiased estimator for Y, $\hat{Y} = \Sigma_i^n w_i Y_i$, is widely known as the Horvitz-Thompson estimator (see *Horvitz-Thompson Estimator**, this volume.) The variance of $\hat{Y}$ is given by

$$\begin{aligned} V(\hat{Y}) &= \Sigma_i^N V(w_i) Y_i^2 + 2\Sigma_i^N \Sigma_{j>i}^N Cov(w_i, w_j) Y_i Y_j \\ &= \Sigma_i^N \pi_i^{-1}(1-\pi_i) Y_i^2 + 2\Sigma_i^N \Sigma_{j>i}^N \pi_i^{-1} \pi_j^{-1} (\pi_{ij} - \pi_i \pi_j) Y_i Y_j \end{aligned}$$

These general results cover a range of the different sample designs described below depending on the values of π_i and π_{ij}. The selection probabilities π_i appear in the estimator and, in addition, the joint selection probabilities π_{ij} appear in the variance. Note that when estimating the parameters of a finite population using the design-based approach for inference, the Y_i values are considered fixed; it is the w_i's that are the random variables.

The selection of a probability sample from a finite population requires the existence of a *sampling frame* for that population. The simplest form of sampling frame is a list of the individual population elements, such as a list of business establishments (when they are the units of analysis). The frame may alternatively be a list of clusters of elements, such as a list of households when the elements are persons. The initial frame may be a list of geographical areas that are sampled at the first stage of selection. These areas are termed *primary sampling units* (PSUs). At the second stage, subareas, or *second stage units*, may be selected within the sampled PSUs, etc. This design, which is known as an *area sample*, is a form of multistage sampling (see below).

The quality of the sampling frame has an important bearing on the quality of the final sample. An ideal sampling frame would contain exactly one listing for each element of the target population and nothing else. Sampling frames used in practice often contain departures from this ideal, in the form of noncoverage, duplicates, clusters, and ineligible units (see Kish 1965, Section 2.7, for a discussion of each of these frame problems.) Issues with the sampling frames used in telephone surveys are discussed in the entry *Telephone Sampling: Frames and Selection Techniques**. Sometimes, two or more sampling frames are used, leading to dual- or multiple-frame designs (see below).

Sampling frames often contain auxiliary information that can be used to improve the efficiency of the survey estimators at the sample design stage, at the estimation stage, or at both stages. Examples are provided below.

2. Simple Random Sampling

A *simple random sample* is a sample design in which every possible sample of size n from the population of N elements has an equal probability of selection

(see *Simple Random Sample**). It may be selected by taking random draws from the set of numbers {1, 2, ..., N}. With simple random sampling, elements have equal probabilities of selection and simple random sampling is therefore an *equal probability selection method (epsem)*.

Simple random sampling with replacement (SRSWR), also known as *unrestricted sampling*, allows population elements to be selected at any draw regardless of their selection on previous draws. Since elements are selected independently with this design, $\pi_{ij}=\pi_i\pi_j$ for all *i, j*. Standard statistical theory and analysis generally assumes SRSWR; this is discussed further in the entry on *Superpopulation Models**.

In *simple random sampling without replacement* (SRSWOR), also simply known as simple random sampling, once an element has been drawn, it is removed from the set of elements eligible for selection on subsequent draws. Since SRSWOR cannot select any element more than once (so that there are n distinct sampled elements), it is more efficient than SRSWR (i.e., the variances of the estimators are lower under SRSWOR than under SRSWR).

3. Systematic Sampling

In the simple case where the *sampling interval* $k=N/n$ is an integer, a *systematic sample* starts with a random selection of one of the first k elements on a list frame, and then selects every k^{th} element thereafter. By randomly sorting the sampling frame, systematic sampling provides a convenient way to select a SRSWOR. Kish (1965, Section 4.1B) describes various techniques for selecting a systematic sample when the sampling interval is not an integer.

If the sampling frame is sorted to place elements that are similar in terms of the survey variables near to each other in the sorted list, then systematic sampling may reduce the variances of the estimates in much the same way as proportionate stratified sampling does. Systematic sampling from such an ordered list is often described as *implicit stratification*. A general drawback to systematic sampling is that the estimation of the variances of survey estimates requires some form of model assumption.

4. Stratified Sampling

Often, the sampling frame contains information that may be used to improve the efficiency of the sample design (i.e., reduce the variances of estimators for a given sample size). *Stratification* involves using information available on the sampling frame to partition the population into L classes, or *strata*, and selecting a sample from each stratum. (See *Stratified Sampling*.)

With *proportionate stratification*, the same sampling fraction (i.e., the ratio of sample size to population size) is used in all the strata, producing an epsem sample design. Proportionate stratification reduces the variances of the survey estimators to the extent that elements within the strata are homogeneous with respect to the survey variables.

With *disproportionate stratification*, different sampling fractions are used in the various strata, leading to a design in which selection probabilities vary. The

unequal selection probabilities are redressed by the use of the base weights in the analysis. One reason for using a disproportionate stratified design is to improve the precision of survey estimates when the element standard deviations differ across the strata. Disproportionate stratified samples are widely used in business surveys for this reason, sampling the larger businesses with greater probabilities, and even taking all of the largest businesses into the sample (see *Business Surveys**). The allocation of a given overall sample size across strata that minimizes the variance of an overall survey estimate is known as *Neyman allocation*. If data collection costs per sampled element differ across strata, it is more efficient to allocate more of the sample to the strata where data collection costs are lower. The sample allocation that maximizes the precision of an overall survey estimate for a given total data collection cost is termed an *optimum allocation*.

A second common reason for using a disproportionate allocation is to produce stratum-level estimates of adequate precision. In this case, smaller strata are often sampled at above average sampling rates in order to generate sufficiently large sample sizes to support the production of separate survey estimates for them.

5. Cluster and Multistage Sampling

In many surveys, it is operationally efficient to sample clusters of population elements rather than to sample the elements directly. One reason is that the sampling frame may be a list that comprises clusters of elements, such as a list of households for a survey of persons (the elements). Another reason is that the population may cover a large geographical area; when the survey data are to be collected by face-to-face interviewing, it is then cost-effective to concentrate the interviews in a sample of areas in order to reduce interviewers' travel. The selection of more than one element in a sampled cluster affects the precision of the survey estimates because elements within the same cluster tend to be similar with respect to many of the variables studied in surveys. The homogeneity of elements within clusters is measured by the *intraclass correlation* (see *Intraclass Correlation Coefficient.*). A positive intraclass correlation decreases the precision of the survey estimates from a cluster sample relative to a SRS with the same number of elements.

When the clusters are small, it is often efficient to include all the population elements in selected clusters, for example, to collect survey data for all persons in sampled households. Such a design is termed a *cluster sample* or more precisely a *single-stage cluster sample* (see *Cluster Sampling**).

Subsampling, or the random selection of elements within clusters, may be used to limit the effect of clustering on the precision of survey estimates. Subsampling is widely used when the clusters are large as, for example, is the case with areal units such as counties or census enumeration districts, schools, and hospitals. A sample design in which a sample of clusters is selected, followed by the selection of a subsample of elements within each sampled cluster is referred to as a *two-stage sample*. *Multistage sampling* is an extension of two-stage sampling, in which there are one or more stages of subsampling of clusters within

the *first-stage units* (or primary sampling units, PSUs) prior to the selection of elements. In multistage sample designs, a key consideration is the determination of the sample size at each stage of selection. This determination is generally based on cost considerations and the contribution of each stage of selection to the variance of the estimator. (See *Multistage Sampling.*)

In general, large clusters vary considerably in the number of elements they contain. Sampling unequal-sized clusters with equal probabilities is inefficient and, with an overall epsem design, it fails to provide control on the sample size. These drawbacks may be addressed by sampling the clusters with *probability proportional to size (PPS) sampling.* By way of illustration, consider a two-stage sample design. At the first stage, clusters are sampled with probabilities proportional to size, where size refers to the number of elements in a cluster. Then, at the second stage, an equal number of population elements is selected within each PSU. The resulting sample is an epsem sample of elements. This approach extends to multi-stage sampling by selecting a PPS sample of clusters at each stage through to the penultimate stage. At the last stage of selection, an equal number of population elements is selected within each cluster sampled at the prior stage of selection. In practice, the exact cluster sizes are rarely known and the procedure is applied with estimated sizes, leading to what is sometimes called *sampling with probability proportional to estimated size* (PPES).

6. Two-Phase Sampling

It would be highly beneficial in some surveys to use certain auxiliary variables for sample design, but those variables are not available on the sampling frame. Similarly, it may be beneficial to use certain auxiliary variables at the estimation stage, but the requisite data for the population are not available. In these cases, *two-phase sampling* (also known as *double sampling*) may be useful. As an example, consider the case where, if frame data were available for certain auxiliary variables, stratification based on these variables with a disproportionate allocation would greatly improve the efficiency of the sample design. Under the two-phase sampling approach, at the first phase, data are collected on the auxiliary variables for a larger preliminary sample. The first-phase sample is then stratified based on the auxiliary variables, and a second phase subsample is selected to obtain the final sample. To be effective, two-phase sampling requires that the first phase data collection can be carried out with little effort or resource requirements.

7. Multiple Frame Sampling

In some cases, more than one sampling frame is available for the target population or part of that population. This situation may arise when one frame provides complete (or nearly complete) coverage of the population but is quite costly to use, while one or more less complete frames provide for less costly sampling and data collection operations. The use of more than one sampling frame is referred to as *multiple frame sampling.* Multiple frame sampling may be

particularly useful for sampling rare populations and in situations in which more than one mode of data collection is needed to adequately cover the population.

An important special case is *dual frame sampling* in which exactly two frames are used. Let A and B denote the two sampling frames, and consider the situation in which frames A and B combined (i.e., $A \cup B$) completely cover the population. Further, partition the population into $a = A \cap B^c$, $b = A^c \cap B$, and $ab = A \cap B$. Note that the population total Y may be expressed as $Y = Y_a + Y_b + Y_{ab}$, where Y_a, Y_b, and Y_{ab} are the totals for subpopulations a, b, and ab, respectively.

With dual frame sampling, the key issue is handling the fact that the elements of the overlap portion, ab, could be selected from either frame. One approach is to eliminate the overlap by identifying elements in the overlap portion with only one of the frames, either prior to sampling or by screening during data collection. An alternative approach is to account for the overlap in estimation, either by computing weights based on the probabilities that the sampled elements could have been selected from either frame or by using the multiple frame methodology originally developed by Hartley (1974). These sampling and estimation approaches are described in detail by Lohr (2009), who also covers more general multiple frame settings.

8. Estimation

As noted above, differential selection probabilities must be accounted for by the use of base weights in estimating the parameters of a finite population. In practice, adjustments are usually made to the base weights to compensate for sample deficiencies and to improve the precision of the survey estimates.

One type of sample deficiency is *unit nonresponse*, or complete lack of response from a sampled element. Compensation for unit nonresponse is typically made by inflating the base weights of similar responding elements in order to also represent the base weights of nonresponding eligible elements (see *Nonresponse in Surveys,* Groves et al. 2001, and Särndal and Lundström 2005).

A second type of deficiency is *noncoverage*, or a failure of the sampling frame to cover some of the elements in the population. Compensation for noncoverage requires population information from an external source. Noncoverage is generally handled through a weighting adjustment using some form of *calibration* adjustment, such as post-stratification (see Särndal 2007). Calibration adjustments also serve to improve the precision of survey estimates that are related to the variables used in calibration.

A third type of deficiency is *item nonresponse*, or the failure to obtain a response to a particular item from a responding element. Item nonresponses are generally accounted for through *imputation*, that is, assigning values for the missing responses (see *Imputation* and Brick and Kalton 1996).

In practice, samples from finite populations are often based on complex designs incorporating stratification, clustering, unequal selection probabilities, systematic sampling, and sometimes, two-phase sampling. The estimation of the variances of the survey estimates needs to take the complex sample design

into account. There are two general methods for estimating variances from complex designs, known as the *Taylor Series* or *linearization* method and the *replication* method (including balanced repeated replications, jackknife repeated replications, and the bootstrap). See Wolter (2007) and Rust and Rao (1996). There are several software programs available for analyzing complex sample survey data using each method.

References

[1] Brick, J.M. and Kalton, G. (1996) Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

[2] Cochran, W.G. (1977). *Sampling Techniques*, 3^nd ed. New York: John Wiley and Sons.

[3] Deming, W.E. (1960). *Sample Design in Business Research*. New York: John Wiley and Sons.

[4] Fuller, W.A. (2009). *Sampling Statistics*. New York: John Wiley and Sons.

[5] Groves, R.M., Dillman, D.A., Eltinge, J.A., and Little, R.J.A., eds. (2001). *Survey Nonresponse*. New York: John Wiley and Sons.

[6] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory*, Vols. I and II. New York: John Wiley and Sons.

[7] Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankyhã, Series C*, 36, 99-118.

[8] Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

[9] Lohr, S.L. (2009). Multiple-frame surveys. In *Handbook of Statistics. Volume 29A, Sample Surveys: Design, Methods and Applications* (Pfeffermann, D., and Rao, C.R., eds.). New York: Elsevier, 71-88.

[10] Lohr, S.L. (2010). *Sampling: Design and Analysis*, 2^nd ed. Pacific Grove, CA: Brooks/Cole.

[11] Murthy, M.N. (1967). *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.

[12] Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), 558-625.

[13] Pfeffermann, D., and Rao, C.R., eds. (2009). *Handbook of Statistics. Volume 29A, Sample Surveys: Design, Methods and Application and Volume 29B, Sample Surveys: Inference and Analysis*. New York: Elsevier.

[14] Raj, D. (1968). *Sampling Theory*. New York: McGraw-Hill.

[15] Rust, K.F. and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

[16] Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99-119.

[17] Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley and Sons.

[18] Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model-assisted Survey Sampling*. New York: Springer-Verlag.

[19] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S, and Asok, C. (1984). *Sampling Theory of Surveys with Applications*, $3^r d$ revised ed. Ames, Iowa and New Delhi: Iowa State University Press and Indian Society of Agricultural Statistics.

[20] Thompson, M.E. (1997). *Theory of Sample Surveys*. London: Chapman and Hall.

[21] Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley and Sons.

[22] Wolter, K. (2007). *Introduction to Variance Estimation*, $2^n d$ ed. New York: Springer.

[23] Yates, F. (1981). *Sampling Methods for Censuses and Surveys*, $4^t h$ ed. London: Charles Griffin.