SAMPLE SURVEYS

S. L. Lohr
Department of Mathematics and Statistics, Arizona State University, USA

Keywords: Cluster sampling, designed experiments, observational data, nonresponse, probability sampling, rare populations, stratified sampling

Contents

Summary

Sample surveys are taken to study characteristics of a population of interest without having to measure every unit in the population. Knowledge about the structure of the population may be used when designing the survey to improve efficiency: commonly used survey designs include simple random sampling, stratified sampling, cluster sampling, unequal probability sampling, systematic sampling, and stratified multistage sampling. The survey design may be used to find standard errors of estimates, which reflect the uncertainty in the estimates due to taking a sample rather than a census. Nonsampling errors such as those caused by nonresponse or measurement error also affect the accuracy of results; although models can be used to adjust for some effects of nonresponse on estimates, it is important to try to minimize these errors when designing the survey.

1. What is a Survey?

Sample surveys are commonly used to provide information about populations of interest in practically every arena of human inquiry. An agricultural survey may be conducted to estimate the number of hectares devoted to different crops in the Republic of Korea in 2001. The primary goal of a public health survey may be estimating the percentage of 5-

year-old children in Nigeria who have been vaccinated against polio, or estimating the prevalence of hypertension in France. Public opinion polls provide snapshots of the public's views on various issues. Social surveys estimate unemployment, income, victimization by crime, poverty, marriage, and other topics. Ecological surveys are used to estimate the densities of wildlife species.

The common feature of all sample surveys is that they study a real, finite, physical population—all hectares of land in Korea, all 5-year-old children in Nigeria, all moths in a region of Honduras, all adults in France—at a specified time. Many surveys are multipurpose in nature, and collect information about a number of responses. The goal of survey design is a sample that provides information about the population with as much accuracy, and as little cost, as possible.

The specificity of the population of interest distinguishes sample surveys from other forms of data collection. The survey's main purpose is generally to estimate characteristics of a specific population at a particular time, not to investigate what could happen under other conditions. The investigator control over units in the population is typically limited to specifying whether they are to be included in the sample. With a designed experiment, however, the investigator can control the behavior of individual units and collect data on what happens if settings are changed.

An example illustrates the difference between a survey and an experiment. An agricultural survey may have the goal of estimating the amount of land that is under cultivation by rice in a specific region in 2001. The investigator can specify which areas are sampled and which measurement methods are used, but typically has no control over whether a unit of land is under cultivation or which variety of rice is planted. It may happen that some land units in the sample are planted with rice variety A and other land units are planted with new, genetically engineered rice variety B. If the average yield from variety B is higher than that for A, however, it is not necessarily true that switching to variety B causes an increase in yield; it could be that farmers who plant variety B engage in other farming practices that increase the yield, or that variety B is typically planted in more productive soil. In contrast, an investigator conducting a designed experiment can randomly assign variety A of rice to some plots and variety B to other plots, and then compare the yields of the two varieties. If variety B has significantly higher yield and if the experiment was conducted properly, the experimenter is justified in concluding that changing the variety likely causes an increase in yield.

A population has N units, where N may or may not be known. The quantity N may be very large, as when the population is all mosquitoes in Illinois, but it is finite. A sample of n units is taken from the N units in the population. The major statistical issues for a survey sample are: what n should be, which units should be selected, and how information from the sample should be used to generalize to the population. Often, information known about the population can be used when designing the survey. For example, in a survey used to investigate employment and income characteristics, the survey designer will know a great deal about different geographic regions: whether they are predominantly urban or rural, their economic status, their ethnic composition, their primary industries, their birth and death rates, and so on. This information can be incorporated into the survey design or estimation methods; generally, the more that is known about the population, the more efficient the survey design can be. The survey design depends heavily on the nature of the population and on the method that is to be used to collect information.

When N is small and collecting data from units is simple, it may be practical to observe every unit—that is, conduct a *census* of the population. In most situations, though, a census, even if possible financially, will not be more accurate than a well-designed and well-implemented survey sample. A census often requires many administrative and field personnel, and the additional complexity may lead to carelessness in data collection or other errors that could be controlled in a smaller sample.

2. Probability Sampling

The primary goal of a survey sample is generally to collect data that can be used to describe features of the population of interest. To achieve this goal, one must be able to use units observed in the sample to make inferences about units in the population that are not in the sample. Results from haphazardly chosen or convenient samples are difficult to generalize to the population because no one knows whether units in the sample can represent units not in the sample. A haphazardly chosen sample often reflects conscious or subconscious biases of the field investigator, who may avoid interviewing persons with large dogs or may be more likely to sample plants closest to the road. Convenience samples often consist of the most accessible members of the population—for example, birds that are easily observed, university students who pass by the library, or computer users who take the time to respond to the internet poll of the day—and it is unknown whether the sampled units are similar to less accessible units in the population. Results from samples that consist entirely of volunteers, such as call-in or e-mail-in polls in which population members decide whether to respond to a broadcast survey announcement, are particularly suspect; since relatively few audience members respond to such surveys, a small group of persons with strong views can determine the survey results.

Some investigators deliberately choose sample units that they believe will be representative of the population units. Such *purposive selection* of a sample may perform well for some measures, but there is no guarantee it will give reliable results. One needs to know the values of the population quantities of interest to be able to quantify how accurately they are estimated by statistics from a purposive sample.

Haphazard, convenient, or purposive samples may provide useful information in some situations. Sometimes no other method of data collection is possible, as when a survey involves intrusive or possibly harmful measurement methods. Even if results cannot be generalized to the population, they may provide useful information; a purposive sample of drinking wells in which 50% of the sampled wells have unacceptable levels of arsenic indicates a public health problem that needs to be addressed. For some surveys, a model may be developed that provides a connection between observed and unobserved units and allows inference within the context of the model. Using a model-based approach to inference, though, requires the untestable assumption that unsampled units are described by the model.

Probability sampling reduces the discretion of the investigator in selecting units, and thus eliminates many of the sources of bias that can occur with haphazard samples. In probability sampling, each subset of the population has a known probability of being included in the sample. A randomizing mechanism such as a random number generator is used to select the sample in accordance with the probabilities, and the probabilities with which different samples could be selected are used to make inferences about the finite population. The probabilities of sample selection are combined with estimates from the sample to construct confidence intervals for population quantities of interest. No assumptions about the joint probability distribution of observations need to be made, and this feature distinguishes inference in survey samples from other areas of statistical inference.

In probability sampling, the probability of inclusion in the sample is known for each unit in the population. Let π_i represent the probability that unit i is included in the sample, and let π_{ij} represent the probability that units i and j are both included in the sample. The sampling *weight* of unit *i* is the reciprocal of the probability with which the unit is selected for the sample. The weight for unit *i* is denoted by $w_i \ (=1/\pi_i)$, and can be interpreted as the number of population units represented by a particular unit in the sample. Let y_i be a characteristic associated with unit *i* in the population. Conceivably, if a census were taken, y_i would be known for every unit in the population; with a sample, y_i is measured only for units in the sample.

The population total, $\sum_{i=1}^{N} y_i$, is often of interest in a survey sample. It is commonly estimated in surveys of all designs by the Horvitz-Thompson estimator, $\sum w_i y_i$, in which the summation is over the units in the sample.

Consider a survey in which 200 people are randomly selected from a population of 10 000 people, and last year's medical expenses (y_i) are determined for every sampled individual. The quantity of interest is the total medical expenses for the population. With this sampling design, every individual in the population has probability $\pi_i =$ 200/10 000 of being included in the sample. The weight of every person in the sample is $w_i = 1/\pi_i = 50$; each person in the sample represents himself/herself plus 49 other persons who were not selected for the sample. The sum of the weights for units in this sample equals the population size, 10 000. Here, calculating the Horvitz-Thompson estimate of the population total is equivalent to creating a new data set of 10 000 "observations" in which each sampled person's response is replicated 50 times, and then summing the values in the new data set. The sampling weights can be thought of as a way of using the sample to construct a "pseudo-population" whose characteristics are thought to resemble those of the original population.

3. Common Probability Sampling Designs

All probability sampling designs require random selection of population units in accordance with the specified selection probabilities. Within that restriction, though, many different designs are possible. The basic designs described below can be used singly or in combination in a survey. The Horvitz-Thompson estimator may be used to estimate the population total with each of these designs. Typically, sampling is done without replacement, so a population unit appears at most once in the sample.

3.1. Simple Random Sampling

In simple random sampling, every possible subset of the population that has size *n* has the same probability of being selected as the sample. Each unit in the population has inclusion probability $\pi_i = n/N$ and sampling weight $w_i = N/n$. The example in Section 2 is a simple random sample of size 200. Simple random sampling is conceptually and analytically simple, but is not always practicable or as cost-effective as some of the more complex designs described below.

3.2. Stratified Sampling

The survey designer can often gain efficiency by dividing the population into distinct groups called *strata*. Often the groups are subpopulations of interest, such as different age or ethnic groups, or different types of terrain. Units within the same stratum are often more homogeneous than units selected randomly from the whole population, so estimates from a stratified sample often have smaller standard errors than could be obtained from a simple random sample of the same size.

A stratified random sample is conducted by taking separate and independent simple random samples from each stratum. In a stratified random sample of 100 men and 100 women from a population of 9 000 men and 1 000 women, each man in the sample has weight $w_i = 90$ and each woman in the sample has weight $w_i = 10$; each sampled man represents himself and 89 other men, and each sampled woman represents herself and 9 other women. The total medical expense for the population of 10 000 people is estimated by [90 × (total expenses for the 100 sampled men) + 10 × (total expenses for the 100 sampled women)]. This stratified design ensures that 100 men and 100 women appear in the sample; with simple random sampling, the number of men will vary from sample to sample.

If every person in the sample has the same weight, the sample design is said to be *self-weighting*. A stratified random sample is self-weighting if the sample size in each stratum is proportional to the population size in that stratum. Self-weighting samples are attractive because they provide a small-scale replica of the population—the sample mean estimates the population mean, the sample median estimates the population median, and graphs of the sample data such as histograms reflect the corresponding graphs that would be constructed from the population if it were known. In many situations, though, non-self-weighting samples better meet the survey objectives. It may be desirable to have larger sampling fractions for minority ethnic groups in a health survey to ensure that the sample sizes are sufficient to provide accurate estimates for each minority group.

-
-
-

Bibliography

Cochran, W.G. (1977). *Sampling Techniques, 3rd ed.* New York: Wiley. [A classic reference on sampling designs and methods of estimation, although it does not contain recent developments in survey sampling.]

Groves, R.M. and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys.* New York: Wiley. [This book discusses effects of and remedies for nonresponse.]

Kalton, G. and Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A*, 149, 65-82.

Levy P.S. and Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications, 3rd ed.* New York: Wiley. [A general introduction to sampling methods, with information on telephone sampling techniques and applications to social and health surveys.]

Lohr S.L. (1999). *Sampling: Design and Analysis.* Pacific Grove CA: Duxbury Press. [This book presents statistical methods underlying sample surveys, emphasizing the importance of survey design and describing resampling methods for variance estimation.]

Rao, J.N.K. (1999). Some current trends in sample survey theory and methods (with discussion). *Sankhyā Series B* **61,** 1-57. [This paper provides insight into important issues in survey sampling at the end of the 20th century, including data collection, inference, and small area estimation.]

Särndal, C.E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag. [The authors use probability models to derive designs and methods of point estimation, but use the sample inclusion probabilities for estimating standard errors.]

Thompson M.E. (1997). *Theory of Sample Surveys.* London: Chapman & Hall. [An advanced book on the mathematical theory underlying sampling methods.]

Thompson, S.K. (2002). *Sampling, 2nd ed.* New York: Wiley. [An introduction to sampling methods, with emphasis on applications to ecological surveys and an exposition of adaptive cluster sampling.]

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* New York: Wiley. [This book uses models to make inferences about finite populations, and describes an alternative to the probability sampling approach given in this article.]

Biographical Sketch

Sharon Lohr was born in Chicago, in the United States; she obtained her B.S. in Mathematics (1982) from Calvin College (U.S.) and her Ph.D. in Statistics (1987) from University of Wisconsin-Madison (U.S.). She worked at the University of Minnesota from 1987 to 1990; since 1990, she has been a professor in the Department of Mathematics and Statistics at Arizona State University. Dr. Lohr's primary research areas are survey sampling, design of experiments, and applications of statistical methods to criminology and law. She is author of the book *Sampling: Design and Analysis*, in addition to refereed journal publications. She has presented several workshops on survey sampling, is a member of the Census Advisory Committee of Professional Associations, and is chair-elect of the Survey Research Methods Section of the American Statistical Association. Dr. Lohr was elected Fellow of the American Statistical Association in 2000, and was chosen as the first recipient of the Gertrude M. Cox Statistics Award in 2003.