Encyclopedia of Survey Research Methods

Probability Sample

Contributors: Erinn M. Hade & Stanley Lemeshow
Editors: Paul J. Lavrakas
Book Title: Encyclopedia of Survey Research Methods
Chapter Title: "Probability Sample"
Pub. Date: 2008
Access Date: December 10, 2014
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412918084
Online ISBN: 9781412963947
DOI: http://dx.doi.org/10.4135/9781412963947.n406
Print pages: 622-624

http://dx.doi.org/10.4135/9781412963947.n406
In gathering data about a group of individuals or items, rather than conducting a full census, very often a sample is taken from a larger population in order to save time and resources. These samples can be classified into two major groups describing the way in which they were chosen: *probability samples* and *nonprob-ability samples*. Both types of samples are made up of a basic unit called an individual, observation, or elementary unit. These are the units whose characteristics are to be measured from a population. In probability samples, each member of the population has a known nonzero probability of being chosen into the sample. By a random process, elements are selected and receive a known probability of being included in the sample; this is not the case in nonprobability sampling.

In order to estimate some quantity of interest with a desired precision from a sample, or to contrast characteristics between groups from a sample, one must rely on knowing to whom or to what population one is referring. Well-designed probability samples ensure that the reference population is known and that selection bias is minimized. The best samples are simply smaller versions of the larger population. The process by which a sample of individuals or items is identified will affect the reliability, validity, and ultimately the accuracy of the estimates and inferences made.

Underlying Concepts

The concepts behind probability sampling underlie statistical theory. From the finite population of *N* elements, all possible samples of size *n* are identified. For example, if the population consisted of 6 elements, and samples of size 2 are to be chosen, there would be 15,

$$\binom{6}{2}$$

, possible samples to consider. In theory, prior to selection, the probability of each of these samples being chosen is known. Therefore, the selection probability of each individual is also known. Summing the probabilities of all samples containing an individual element will compute the individual's probability of appearing in the sample. By knowing the probability of selecting each unit, a statistical weight can be assigned by which population estimates can be calculated. The statistical weight is defined as the inverse of the probability of selection into the sample, allowing each sampled unit to represent a certain number of units in the population. Most often, the goal of a probability sample is to estimate certain quantities in the population using these statistical weights.

Reliability and Validity

From a probability sample, the quantities that are estimated, such as population totals, means, proportions, and variances, have certain properties that can be evaluated. For instance, over repeated sampling, estimators from the probability sample can be evaluated for how reproducible or reliable they are (variance) and if, on average, the estimates from the sample are similar to the true value of the population quantity (unbiasedness or validity). Combining the ideas of reliability and validity, the accuracy, or how far away on average the estimator is from the true value (mean squared error), can also be evaluated on the sample estimator.

None of these desirable properties can be determined from estimates derived from nonprobability samples. Nonprobability samples are used in many unscientific surveys, market research, and public opinion polls, often because they are easier and less expensive to conduct. These types of surveys include purposive or deliberate, quota, and snowball samples. As an example, imagine that interviewers are attempting to question shoppers as to their political views at a local supermarket in order to describe future poll results for a city. In quota sampling, the interviewer may be asked to “find” a certain number of individuals in various demographic groups, such young women, older women, black men, and white men. The individuals that are found by the interviewer may be of only one political leaning or of **[p. 622 ↓]** one socioeconomic group simply because they are easy to find and shop at the particular market. Without a systematic plan for sampling, the decision about whom to interview is left up to the interviewer, likely creating bias in the sample. In certain circumstances, such as in populations that are hard to reach, probability samples may not be feasible. Thus, as long as they are not used to make inference to a larger population, some non-probability samples are useful.

In order to draw a probability sample, a list of sampling units must be organized from which to sample. In hard-to-reach populations, this sampling frame may not be available. The sampling frame must allow every unit in the population to have a chance of selection; otherwise coverage error could result. For example, the sampling frame may consist of all medical records within a certain time period. A sample can be drawn from the medical record numbers. Alternatively, the sampling frame may be an enumerated list of city blocks from which a household, and eventually an individual, will be chosen. Using the sampling frame, the sample can be selected in numerous ways.

Sampling Designs

Simple Random Sampling

The simplest design used to sample units is called a “simple random sample.” Simple random sampling can either be done *with replacement* or *without replacement*. In sampling with replacement, the unit sampled can be selected more than once, since it is returned to the pool once sampled. In practice, nearly all probability samples are performed without replacement, having each element appear only once in the sample. Simple random samples without replacement consist of selecting n units from a population of N elements, each possible subset having the same probability of selection. For example, if numbers 1 to 100 were placed in an urn, and 10 numbers were drawn without replacing them after each turn, this would be a simple random sample. The units or people associated with these 10 numbers would constitute the sample from the population of size 100.

In fact, simple random samples are not often used alone as the sample design for a survey. Enumerating every element of a population is a tedious and time-consuming process. In addition, once this list is compiled it may be out of date. Perhaps a study's interest is in estimating a certain health characteristic of high school students in a city. If one were to enumerate all high school students in the city today, students may leave or enter the school tomorrow. Moreover, performing a simple random sample design for a population that covers a large geographic area is not practical. Since each element will have the sample probability of selection, it may require traveling many miles between elements. For all of these reasons, simple random sampling is rarely used alone; however, it provides the basis for comparison to other more commonly used designs.

Systematic Sampling

In some situations, it is not possible to construct a sampling frame before sampling has to occur, but a very good estimate of the number of records in a certain interval may be known. In situations like these it is possible to take a probability sample comprised of every *kth* element in the population. This is called “systematic sampling.” Suppose characteristics of patients presenting to an emergency room is being planned and, from previous observation, it is known that 1,000 patients will come to the emergency room in a given week. If we would like to collect information on 100 patients using a systematic sample, we will survey every 10th patient. A random number is selected between 1 and 10, and every 10th element after that is included in the sample. Systematic sampling is an extremely popular sampling method due to its ease of implementation.

Stratified Random Sampling

One way to improve the precision of estimates over what is possible with simple random sampling is to carry out what is called a “stratified random sample.” In this type of sample, elements of a population may also be categorized into distinct groups, called “strata.” Within each stratum, a sample of units can then be drawn using simple random or systematic sampling. In general, stratification by a characteristic will reduce variability in the resulting population estimates, especially when the characteristic is related to the measurement of interest. Often, it will also allow reliable estimates to be made about each stratum. Following from the previous example, the individuals to sample may be either male or female. Separately, a simple random sample of men and a simple random sample of women can be chosen, stratifying the sample by gender. Disadvantages of stratified random **[p. 623 ↓]** sampling still exist, including the cost of constructing a sampling frame for each stratum before drawing the sample and then a probability sample from each. This may actually increase the costs of selecting the sample over what would be the case for a simple random sample, but the increased precision may justify the additional time and cost.

Cluster Sampling

One common variation in designing a probability sample is called "cluster sampling." In this type of sampling, the sampling frame enumerates listing units that are not the individuals or elements, but are larger groupings called "clusters." For instance, it may be easier to enumerate city blocks and sample these blocks rather than enumerate all individuals to conduct a survey describing the health of individuals in an area. Once blocks are chosen, dwelling units can be sampled and individuals selected within each unit. Cluster sampling may dramatically reduce cost and time; however, these designs usually have a trade-off in precision. As opposed to simple random sampling, sampling frames are much easier to construct for cluster samples. Only the elementary units of the chosen clusters will have to be listed for sampling. This may entail counting the number of houses on a city block or students in a particular classroom as opposed to the entire city or school. Although clusters may be chosen randomly, most often they are chosen with probability proportional to some measure of the cluster's size (PPS). In general, cluster sampling with PPS reduces variability in estimates as compared to cluster sampling with equal probability, when the size of clusters varies greatly. Moreover, since cluster samples greatly decrease travel costs, the decreased precision for a fixed sample size compared to simple random sampling is outweighed by the ability to sample a greater number of individuals for a fixed cost, ultimately resulting in more precise estimates for a fixed budget. Cluster sampling is feasible in many situations where simple random sampling is not.

Some common *random-digit dialing* (RDD) techniques are, in fact, cluster samples. The most straightforward random-digit dialing method simply randomly selects phone numbers from a frame: a simple random sample. However, only 20% of the phone numbers may be household phone numbers. RDD methods based on sampling combinations of the area code, the exchange, and blocks of the remaining four numbers are cluster samples. *Blocks* are groups of 100 numbers with the same first two digits. For example, 555-444-12XX would be considered one 100-block listed on the sampling frame. Once a household is found in an area code + exchange + block, all of the numbers in the particular block are called, dramatically reducing the number of nonhousehold calls, as is done in the Mitofsky-Waksberg approach to RDD sampling.

Importance of Well-Designed Probability Surveys

The elements of design, including stratification, clustering, and statistical weights, should not be ignored in analyzing the results of a probability sample survey. Ignoring the sampling design may underestimate variability, affecting potential inferences. Software has advanced greatly in recent years and has become more accessible to researchers wishing to estimate population characteristics and their associated standard errors using sample survey data. This should encourage researchers to carry out well-designed probability surveys to attain their study objectives and to use appropriate methods to analyze their data.

Erinn M.Hade, and StanleyLemeshow

http://dx.doi.org/10.4135/9781412963947.n406
See also

- Census
- Cluster Sample
- Coverage Error
- Elements
- Finite Population
- Mitosky-Waksberg Sampling
- Multi-Stage Sample
- n
- N
- Nonprobability Sample
- Probability Proportional to Size (PPS) Sampling
- Purposive Sample
- Quota Sampling
- Random

- Random-Digit Dialing (RDD)
- Random Sampling
- Replacement
- Sampling Frame
- Simple Random Sample
- Snowball Sampling
- Strata
- Stratified Sampling
- Systematic Sampling
- Target Population
- Unbiased Statistic
- Unit of Observation
- Validity
- Variance

Further Readings

Levy, P. S., & Lemeshow, S. (1999). Sampling of populations: Methods and applications (3rd ed.). New York: Wiley.

Lohr, S. L. (1999). Sampling: Design and analysis . Pacific Grove, CA: Duxbury.