Random Sampling

In: Encyclopedia of Research Design

By: Mosuk Chow

Edited by: Neil J. Salkind
Book Title: Encyclopedia of Research Design
Chapter Title: "Random Sampling"
Pub. Date: 2012
Access Date: August 28, 2017
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412961271
Online ISBN: 9781412961288
DOI: http://dx.doi.org/10.4135/9781412961288
Print pages: 1211-1213

In random sampling, the sample is drawn according to prespecified chances from the population, and thus it is also called probability sampling. Since planned randomness is built into the sampling design according to the probabilities, one can use these probabilities to make inferences about the population. For example, if one uses the sample mean to estimate the population mean, it is important to know how the sample is being drawn since the inference procedures such as confidence intervals will depend on the sampling scheme. Similarly, hypothesis testing on the population mean also depends on the sampling scheme used.

One important purpose of random sampling is to draw inferences about the population. On the other hand, if the sampling scheme is nonrandom, that is, not all outcomes have a known chance of occurring, the sample is likely to be biased. It is difficult to draw inferences about the population on the basis of nonrandom sampling. Some common nonrandom sampling methods are quota sampling, convenience sampling, and volunteer sampling.

The following sections provide a brief outline of some of the most commonly used random sampling schemes: simple random sampling, systematic sampling, stratified sampling, and cluster sampling.

Simple Random Sampling

Simple random sampling is the simplest random sampling scheme. There are two types of simple random sampling: with or without replacement. For simple random sampling without replacement, (n distinct units are selected from a population of N units so that every possible sample has the same chance of being selected). Thus the probability of selecting any individual sample s of n units is

$$p(s) = \frac{1}{\binom{N}{n}} \quad \text{where} \quad \binom{N}{n} = \frac{n!(N-n)!}{N!}.$$

Sampling without replacement is often preferred because it provides an estimate with smaller variance. However, even though for a finite population, when a sample has been selected, it does not provide additional information to be included in the sample again. One practical advantage to sample with replacement is there that is no need to determine whether a unit has already been sampled. In addition, when sampling with replacement, each drawing of each sample is independent, and the probability formulas for the samples are usually easier to derive. For simple random sample with replacement, a unit may be selected more than once, and each draw of a unit is independent: Each unit in the population has the same probability of being included in the sample. When the sample size is much smaller than the population size, the two sampling schemes are about the same. Otherwise, it is important to distinguish whether the simple random sampling is with or without replacement since the variance for statistics using these two schemes is different, and thus most inference procedures for these two schemes will be different. For example, in estimating the population mean under simple random sampling without replacement, the estimated variance for the sample mean is

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

where s2 is the sample variance.

Note that the factor

$$\left(1 - \frac{n}{N}\right)$$

is called the finite population correction factor. Under simple random sampling with replacement, there is no need for this factor, and the estimated variance for the sample mean is simply

$$\hat{V}(\bar{y}) = \frac{s^2}{n}.$$

In designing a survey, one important question is how many to sample. The answer depends on the inference question one wants to answer. If one wants to obtain a confidence interval for the parameter of interest, then one can specify the width and level of significance of the confidence interval. Thompson (2002), chapter 4, provides an excellent discussion for determining sample sizes when one is using a simple random sample.

One advantage of simple random sampling is that making inferences is simple and easy with this sampling scheme. It is easy to apply to small populations. With a large population, it is often difficult for researchers to list all items before they draw randomly from the list. This difficulty limits the use of simple random sampling with large populations.

Systematic Sampling

Systematic sampling is easier to perform in the field and provides a useful alternative to simple random sampling. The researcher starts at a random point and selects items that are a fixed distance apart. When no list of the population exists or if a list is in approximately random order, systematic sampling is often used as a proxy for simple random sampling. In cases in which the population is listed in a monotone order, systematic sampling usually results in estimators with smaller (though sometimes unestimable) variances than those of simple random sampling. Repeated systematic sampling consists of more than one systematic sample, each with a different random starting point. Using the variability in the subsample means, one can get a measure of the variance of that estimate in the entire sample. When the population is in periodic order, then it is important to choose the period appropriately in order to get a representative sample.

One can classify systematic sampling as a special case of cluster sampling. The population is partitioned into clusters such that the clusters consist of units that are not contiguous. For systematic sampling to be effective, the ideal cluster should contain the full diversity of the population and thus be representative of the population. One simple example for systematic sampling is to interview subjects in a long queue about their political affiliation. If there are 700 people in the queue and one plans to take a sample of size 50, then to take a systematic sample, one chooses a number randomly between 1 and 14 and then every 14th element from that number on. If the people form the queue in a random order, then systematic sampling will produce the same result as a simple random sample. If people with similar political affiliation stay close together in the queue, then systematic sampling will likely be more effective than simple random sampling.

Systematic sampling is often used in industry for quality control purposes. When there is a malfunction beginning with and continuing from a certain item, a systematic sample will provide a representative sample of the population whereas simple random sampling may over- or underrepresent the defective items.

Stratified Sampling

When the population is partitioned into strata and a sample is selected from each stratum, then the sampling scheme is called stratified sampling. One important characteristic of the strata is that the elements in the population belong to one and only one stratum. Stratified random sampling refers to the situation in which the sample selected from each stratum is selected by simple random sampling. Alternatively, one may select a sample from each stratum by systematic sampling or other schemes. When the stratum consists of units more similar in the variable of interest, stratified sampling usually results in estimates with smaller variances. One example of stratified sampling is a study to estimate the average number of hours per week that undergraduates from a large public university spend studying. One can stratify by department and conduct stratified sampling. If the time that students within a department spend studying is about the same whereas there is large variability between departments, then a stratified sample will have smaller variance than a random sample will. On the other hand, if one wants to estimate the average height of undergraduate students in that university, one may choose to use stratified random sampling by department for ease of administration. Note that the resulting estimate from stratified sampling may not have the benefit of smaller variances than simple random sampling would offer because students in the same department may be of very different height.

One important practical consideration in stratified sampling is how many samples to allocate to each stratum. When one has no prior information about the variances of the strata, then the optimal allocation is to assign sample sizes proportional to stratum size. If the sampling cost is the same for each stratum and the standard deviations of each stratum can be estimated, then the optimal allocation is to assign sample sizes proportional to the product of stratum size and stratum standard deviation. That is, the optimal allocation assigns larger sample sizes to larger or more variable strata.

Quite often, the characteristic that one wants to base the stratification on may not be known before the survey is carried out. For example, we may want to stratify voters by gender, but gender information may be known only after the voter is contacted. Double sampling (also called two-phase sampling) may be used in such situations. In double sampling, a simple random sample is collected from the population for the purpose of classifying these items to the appropriate stratum. A second sample is then selected by stratified random sampling from the first sample. One important use of double sampling is to adjust for nonresponse in surveys. In addition to the responses they obtain, the researchers stratify the population into response and nonresponse groups according to a first sample of the population. To obtain responses, the researchers re-calls people from the nonresponse group and offer them more incentives to respond. See Scheaffer, Mendenhall, and Ott (1995), Chapter 11.5, for guidelines for determining the number of callbacks.

Cluster Sampling

Sometimes listing the elements of a population is costly or not available, but a listing of clusters is available. Also, the cost of sampling elements within a cluster is often less than the cost of other sampling methods. In such cases, cluster sampling is a good choice. Cluster sampling and stratified sampling differ in that in stratified sampling, a simple random sample is taken from every stratum, whereas in cluster sampling, a simple random sampling of the clusters is taken. The effectiveness of cluster sampling depends on the variances from using the sampled clusters and the costs of sampling these clusters and the units within the clusters. Quite often, cluster sampling has less precision than simple random sampling from the same universe because units within the same cluster are usually more similar and thus provide less information per unit. In cluster sampling (or one-stage cluster sampling), all units of a cluster are sampled if the cluster is a sample. If, after selecting a sample of clusters, a sample of units within the clusters is selected, the design is called two-stage cluster sampling. Similarly, multistage cluster sampling may be carried out. Suppose one wants to estimate the average hours per week undergraduates from a large public university spend studying. One can treat each university department as a cluster. To conduct one-stage cluster sampling, one randomly selects departments in that university, then samples all the undergraduate students in those departments. If instead one samples some individuals from the selected departments, one is using two-stage cluster sampling. An example for multistage sampling is to first select states, then universities within the states, then departments within the universities, and then students within the departments.

In two-stage cluster sampling, various designs exist for selecting the clusters in the first stage. Frequently one uses probability proportional to size, in which the probability of a cluster's being selected is proportional to the number of units it contains. Other sampling schemes commonly employed at each stage include simple random sampling and systematic sampling.

When designing a cluster sample, one decides what overall precision is needed, what size the clusters should be, and how many to sample from each cluster. Lohr (1999) Chapter 5 provides a detailed discussion.

Conclusion

For a complex survey, rarely will only one type of sampling scheme be used. Often many sampling schemes will be combined. For example, a population may be divided into strata, and multistage cluster sampling can be used to sample from each stratum. Simple random sampling may be used in one stage whereas systematic sampling be used in the next stage. The four basic random sampling schemes discussed here have important implications for the estimators to use and the inference procedure for each estimator.

MosukChow

http://dx.doi.org/10.4135/9781412961288.n364

See also

- Population
- Sample
- Sampling

Further Readings

Cochran, W. G.(1977). Sampling techniques (3rd ed.). New York: Wiley.

Kish, L.(1965). Survey sampling.New York: Wiley.

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

Scheaffer, R. L.,Mendenhall, W., &Ott, R. L.(1995). Elementary survey sampling (5th ed.). Pacific Grove, CA: Duxbury.

Thompson, S. K.(2002). Sampling (2nd ed.). New York: Wiley.