Probability Sampling

In: Encyclopedia of Research Design

By: David L. R. Affleck
Edited by: Neil J. Salkind
Book Title: Encyclopedia of Research Design
Chapter Title: "Probability 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: 1110-1112

In many scientific inquiries, it is impossible to individually appraise all the elements that comprise a population of interest. Instead, there is a need to infer collective properties of the whole population from a select subset of its elements. Probability sampling is an approach to selecting elements from a fixed population in such a way that

- elements are selected by a random process,
- every element has a nonzero chance of selection, and
- the relative frequency with which an element is included in a sample is deducible.

A collection of elements drawn in such a way is referred to as a probability sample. The first condition imparts a degree of objectivity to a probability sample and, in combination with the other two conditions, secures a basis for statistical inferences concerning descriptive parameters of the population. It is also largely because its integrity rests on these conditions of the selection process, rather than on the acuity of the investigator, that probability sampling is so widely adopted in modern scientific and statistical surveys.

Sample Selection

In probability sampling, the selection probabilities of individual population elements and the algorithm with which these are randomly selected are specified by a sampling design. In turn, to apply a sampling design requires a device or frame that delineates the extent of the population of interest. A population often can be sampled directly using a list frame that identifies all the elements in that population, as when the names of all students in a district are registered on school records. If a list frame is available, a probability sample can be formed as each of a string of random numbers generated in accordance with a design is matched to an element or cluster of elements in the population. By contrast, some populations can be framed only by their spatial boundaries. For example, many natural resource populations can be delineated only by the tracts of land over (or under) which they are dispersed. Such a population must be surveyed indirectly via a probability sampling of coordinate locations within its spatial domain. Regardless of how a population is framed, however, the frame must be complete in the sense that it includes the entirety of the population. Inasmuch as any fraction of the population omitted from the frame will have zero probability of being selected, the frame ultimately fixes the population to which probability sampling inferences apply.

In some applications, once a design is chosen, the set of all selectable probability samples can be discerned together with the relative frequency with which each sample will be drawn. In other cases, the size of the population is never known and one cannot calculate even the number of possible samples. Nevertheless, when a valid probability sampling design is

employed, at a minimum it is possible to derive the inclusion probabilities of the elements ultimately selected. The inclusion probability of an element is the relative frequency with which it is included in the observation set. Some probability sampling designs prescribe equal inclusion probabilities for all elements, but most allow for variation across the population. Designs of the latter variety intrinsically favor the selection of certain elements or classes of elements, but only to an extent that can be discerned from the relative magnitudes of the elements' inclusion probabilities.

Notably, even where a design skews the inclusion probabilities toward a certain class of elements, random selection precludes both the investigator (and the elements themselves) from directly influencing the composition of a probability sample. To underscore this point, it is informative to contrast probability sampling with non-probability sampling methods, such as purposive or quota sampling. In these strategies, the investigator plays a direct role in the selection process, often with the aim of assembling a sample that is in some sense representative or typical of the population. By definition, population elements are not drawn objectively, and the likelihood of having observed one element as opposed to another is inscrutable. A non-probability sample can yield an estimate that is close in value to a particular population parameter, but because of its discretional nature, the selection process provides no basis for assessing the potential error of that estimate. In counterpoint, probability sampling is not designed to select representative samples, save in the weak sense of allowing every element a nonzero probability of being observed. Probability sampling is instead fomulated to objectify the selection process so as to permit valid assessments of the distribution of sample-based estimates.

Estimation and inference

Applied to a given population, probability sampling admits the selection of any one of a typically large number of possible samples. As a result, a chosen parameter estimator acquires a distribution of possible values, each value corresponding to at least one selectable sample. This distribution is often referred to as the randomization distribution of an estimator because it arises from the random selection process. In practice, an estimator takes only one value when applied to the sample data actually collected, but this singular estimate is nonetheless a realization of the randomization distribution induced by the sample design.

The properties of an estimator's randomization distribution are conditioned by various aspects of the population being sampled but are also controlled by the sampling design. Indeed, an advantage of probability sampling is that the character and degree of variability in an estimator's randomization distribution often can be derived from the sampling design and the

estimator's algebraic form. In particular, without setting conditions on the structure of the population, one can generally specify estimators that are unbiased for specific parameters. For example, take the Horvitz–Thompson estimator,

$$\hat{\theta} = \sum_{k \in s} \frac{y_k}{\pi_k},$$

which expands the attribute measurement yk made on the kth distinct element in a probability sample s by that element's inclusion probability (πk) and sums these expansions across all selected elements. This estimator is unbiased for the true population total of the yk under any probability sampling design. Thus, to unbiasedly estimate total research expenditures in a particular sector of the economy, a probability sample of firms in that sector could be drawn and then the sum of the ratios of each firm's research outlay (yk) to inclusion probability (πk) could be calculated.

Modifications of the Horvitz–Thompson and other estimation rules provide unbiased estimators of population means, proportions, and distribution functions. Additionally, for many designs, unbiased estimators of the variance of these parameter estimators can be derived. Hence, with many probability sampling strategies, one can estimate population parameters of interest in an unbiased manner and attach valid measures of precision to those estimates. Going further, one can draw inferences on parameters via confidence intervals that account for the variability of the estimator, given the sampling design.

Designs

The random selection of elements in probability sampling introduces the possibility of obtaining estimates that deviate markedly from the parameter(s) of interest. Fortunately, there exist numerous design features that dampen the magnitude of sample-to-sample variation in estimates and reduce the likelihood of large deviations. Broadly, the accuracy of an estimator can be improved by applying designs that enhance the dispersion of sample elements across the population or that exploit pertinent auxiliary information on the structure of that population.

Simple random sampling (SRS) is among the most basic of probability sampling designs. From a list frame, a fixed number n of independent selections are made with all elements having identical inclusion probabilities. Whether elements are selected with or without replacement, SRS results in every combination of n elements being drawn with the same frequency. Without replacement, SRS carries the advantage that, on average, individual samples contain a larger number of distinct population elements. Therefore, more information about the population is collected than when element-replacement is permitted, and there is less sample-to-sample

variation among parameter estimates.

By eliminating the independence of selections, systematic sampling provides a direct means of drawing dispersed sets of elements. With a systematic probability sampling design, one element is randomly selected from the frame, and then all elements that are separated from this initial selection by a fixed sampling interval are added to the sample. A sample of citizens voting at one location, for example, could be drawn by rolling a die to select one of the first six voters and then subsequently interviewing every sixth voter to exit the polling station. Systematic sampling generally ensures that all elements have equal inclusion probabilities but, at the same time, renders observable only those combinations of elements that are congruent with the sampling interval. Systematic designs are particularly efficient when the sampling interval separates selected elements along a population gradient, be it a natural gradient or one artificially imposed by ordering the frame. By ensuring that every possible sample spans such a gradient, a systematic design reduces variation among estimates and generally improves precision relative to SRS. Unfortunately, because sample elements are not selected independently, it is difficult to assess the precision of estimates if only a single systematic sample is collected.

If prior information exists on natural classifications of the elements of a population, a stratified sampling design can be employed. Such a design divides the population into an exhaustive set of disjoint strata and specifies the collection of mutually independent probability samples from each. Thus, pupils in a school district or trees in a forest stand might be stratified into elementary and secondary students, or into conifer and deciduous species, prior to sampling. As sample elements are necessarily drawn from each stratum, stratified designs permit estimation of stratum-level parameters while ensuring a broad level of coverage across the population as a whole. Moreover, if a significant proportion of the variability in the attribute of interest (e.g., student attendance rate or tree biomass) is due to differences among strata, stratification can lead to large gains in precision relative to SRS.

Stratified sampling designs often vary the inclusion probabilities across strata in order to sample larger, more variable, or more important strata with higher intensity. This concept is carried further by many unequal probability sampling designs, such as Poisson sampling and list sampling. These designs are most effective when the inclusion probability of each element can be made approximately proportional to the magnitude of the attribute of interest. Often this is achieved by making the inclusion probabilities proportional to a readily available auxiliary variable that is in turn positively correlated with the attribute of interest. Thus, if interest centers on total research expenditures in a manufacturing sector, firms with a larger number of employees might be assigned larger inclusion probabilities; if total tree biomass in a stand is of

interest, trees with large basal diameters might be assigned larger inclusion probabilities. The attribute measurements taken on the sampled elements can be weighted by their respective inclusion probabilities to secure unbiased estimation. Additionally, if the inclusion probabilities have been judiciously chosen, the probability-weighted attribute measurements can be appreciably less variable than the raw attribute scores, which can lead to substantially less variation among estimates than is seen under SRS.

One area of active research in probability sampling is the incorporation of statistical models into sample selection and estimation strategies. In many cases, this offers the potential to improve accuracy without sacrificing the objectivity of the probability sampling design as the basis for inference. Of course, in many applications, statistical models can be used to great effect as a basis for inference, but the validity of inferences so drawn then rest on the veracity of the presumed model rather than on the sample selection process itself.

David L. R.Affleck

http://dx.doi.org/10.4135/9781412961288.n337

See also

- Estimation
- Nonprobability Sampling
- Parameters
- Population
- Random Sampling

Further Readings

Gregoire, T. G., &Valentine, H. T.(2008). Sampling strategies for natural resources and the environment.Boca Raton, FL: Chapman & Hall/CRC.

Overton, W. S. , & Stehman, S. V. *The Horvitz-Thompson theorem as a unifying perspective for probability sampling: With examples from natural resource sampling.* (1995).49,261–268.

Sarndal, C.-E.,Swensson, B., &Wretman, J.(1992). Model assisted survey sampling.New York: Springer.