ANNOTATED GLOSSARY

OF

SELECTED SAMPLING TERMS

Earl E. Houseman

U.S. Department of Agriculture
Economics, Statistics, and Cooperatives Service
Washington, D.C. 20250

JULY 1979

ANNOTATED GLOSSARY OF SELECTED SAMPLING TERMS. By Earl E. Houseman; Statistical Consultant, Retired from U.S. Department of Agriculture.

ABSTRACT

Important statistical terms commonly used in sample surveys are defined and discussed briefly. Definitions of such terms, as found in the literature on mathematical statistics, are often inadequate in practice. This glossary is more than concise definitions of terms. Explanations of important concepts associated with the terms are included with regard to practical conditions encountered in surveys.

Key words: glossary; sampling terms; sampling error; bias; surveys.

Preface

Clear concepts and complete definitions are an essential part of good statistical practice. This annotated glossary contains a selection of sampling terms that pertain to probability sample surveys. No attempt was made to be exhaustive in the selection of terms. Instead, attention is focused on definitions that involve important concepts, philosophies, or principles in probability sample surveys. The glossary is annotated to expand the meaning of terms without getting detailed.

An effort has been made to be consistent with common usage. However, it is important to remember that statisticians and organizations tend to coin their own definitions. Choice of terms and details of definitions must be worked out for specific cases as needed.

Because the defintions are related to one another, they are not listed in alphabetical order. The order is intended as a logical order if one were to study the definitions from beginning to end. An alphabetic listing of the terms is appended.

1. Statistical Survey - an investigation involving the collection of data. Observations or measurements are taken on a sample of elements for making inferences about a defined group of elements. Surveys are conducted in many ways. The basis for the following discussion of terms is sample surveys in which probability sampling is employed.

2. Population - a fully defined set of elements that a survey pertains to. That is, from a sample inferences are made about a set of elements called a population. Complete specifications of the elements which comprise the population is a necessary step in the application of probability sampling, although it is not always necessary to have every element identified and listed prior to the selection of a probability sample. The elements of the population must be defined in space and time as well as content. For example, after fully defining a farm, the population to be surveyed might be described as all farms in the State of Texas that were in operation during August 1970. In this case, "in operation" would also need to be defined.

For a variety of practical reasons, the definitions of a population for a survey might differ from one which is preferred in conceptual or idealistic terms. For example, farms in remote areas where only a few widely scattered small farms are located might not be included owing to cost.

3. Universe - often used synonymously with population. "Population" refers either to the elements of a set or to a set of data for some characteristic of the elements. "Universe" usually refers only to the elements.

4. Characteristic - a general term for any variable or attribute. Observations or measurements are taken on specified characteristics of elements in a sample.

5. Unit of Observation - a unit for which data are obtained.

6. Unit of Analysis - a unit for tabulation or analytical purposes. In many cases, "element," "unit of observation," and "unit of analysis" are the same unit.

7. Statistical Population - a set of data for a characteristic of all elements of a population. Assuming N elements in the population, a typical representation of the N values of the characteristic X is $X_1, X_2, \ldots, X_N$ and X_i represents the value of X for the i-th element. The word "population" is often used in lieu of "statistical population" when referring to data.

8. Probability Sample - a sample obtained by application of the theory of probability. In probability sampling, every element in a defined population has a known, nonzero, mathematical probability of being selected. It should be possible to consider any element of the population and state its probability of selection.

9. Sampling Unit - a unit of the population that may be selected when one random selection is made for the sample. The sampling units might be individual elements of the population or groups of elements. Some sampling units might not contain an element. If each element in the population belongs to one and only one sampling unit, the aggregate of all sampling units includes all elements of the population without omission or duplication. In complex situations, an element might not be uniquely associated with only one sampling unit; but it should have an ascertainable probability of being associated with any given sampling unit.

The "sampling unit" plays a central part in sampling theory. Sampling variance of an estimate is a function of variation among sampling units. "Sampling units" refers to all sampling units in the population as well as the units that happen to be in a sample. The term "sample units" could apply to the units selected for a sample.

10. Simple Random Sampling - a special case of probability sampling. Sometimes called unrestricted random sampling. It is a process for selecting n sampling units one at a time, from a population of N sampling units so each sampling unit has an equal chance of being in the sample. Every possible combination of n sampling units has the same chance of being chosen. Selecting one sampling unit at a time with equal probability, without replacement, using tables of random numbers satisfies this definition of simple random sampling.

11. Sample Design - specifications for selecting a sample. Such specifications are determined with regard to the purpose of the survey and achieving low sampling errors (efficient sampling for a given cost). Simple random sampling as defined above, is seldom used in practice but the theory for it is of fundamental importance and extends readily to many sampling plans (for example, stratified random sampling). Reducing sampling error by improving the sample design is often less expensive than by increasing the size of the sample.

12. Census - a statistical survey which, from a mathematical viewpoint, is a special case of sampling. That is, sampling is the general case and a census is the special case when all elements of the population are assigned a probability equal to one of being included in the sample. There are many sources of error which pertain to results from a census survey, as well as from a sample survey.

13. Statistics - totals, averages, percentages, and other numbers computed from population or sample data. "Statistics" also means statistical methods.

14. Sampling Frame - a list, or the equivalent of a list, of all sampling units of the population. It enables probability sampling. The phrase "or the equivalent of a list" is inserted because frames often account for all sampling units (and elements) in a population without having every sampling

unit explicitly listed or defined in advance. A frame is needed whether sampling or a census is involved.

15. Frame Units - units of a sampling frame. They might be identical to or different from elements of the population or from sampling units. Cards in a file, blocks of a city, and classes in a university are examples of frame units. Frame units are often subdivided or combined to form sampling units.

16. Probability of Inclusion - the probability which any given sampling unit in the population has of being included in a sample. For example, with a population of N sampling units, each sampling unit in the population has a probability equal to $\frac{n}{N}$ of being in a simple random sample of n units. An element has the same probability of inclusion as the sampling unit which it belongs to unless the sampling unit is subsampled.

17. Probability of Selection - the probability of a sampling unit being selected at any given random draw. Although "probability of selection" is often used as meaning "probability of inclusion," the probability that a unit has of being selected at a given random draw and of being included in the sample is an important distinction.

18. Population Parameters - unknown quantities to be estimated. A parameter is defined with reference to a set of unknown values of X for the N elements of the population. Quantities derived from this set without any sampling would be parameters.

It is important to recognize that the value of each X_i in the set is subject to possible variation attributable to details of the definitions of the characteristics, methods of measurement, and various operating conditions. For most characteristics, one can readily identify factors that could influence the value of X_i. For example, if X_i is the weight of the i-th person, the value of X_i that gets recorded on a questionnaire depends on such things as who answered the question and how and when the weight was determined. If X_i is the acreage of a field of corn, its value depends on details of the definition of corn acreage and how the acreage is determined under any given definition. Fortunately, small variation in the measured value of X_i for the i-th element is unimportant for most statistical purposes; but, sometimes small differences are of critical importance. Parameters should not be thought of as invariant under procedural changes even though the definition remains unchanged because each X_i is subject to some degree of variation or error which depends on the conditions involved in determining a value for it.

19. True Values - strictly speaking, are conceptual or hypothetical because measurement error exists to some degree, even if negligible for practical

purposes. In some situations, it might be helpful to assume a set of true values $T_1,\ldots, T_N$ for a characteristic. Then, for example, the average value of T_i would be the true average. True values can only be approximated with increasing closeness by improving measurement methods. Measurements $X_1,\ldots,X_N$ for the N elements of the population may be called population values. Likewise, values of X in a sample, $x_1,\ldots,x_n$, may be called sample values. The term true values is often used loosely.

20. Random Variables - a variable which, by chance, can be equal to any value in a specified set. The probability that it equals any given value (or falls between two limits) is either known, can be determined, or can be approximated or estimated. A chance mechanism determines the value which a random variable equals. A function of random variables is also a random variable. For example, the value of x_i for the i-th element in a probability sample is a random variable because the element was randomly selected. An estimate based on a probability sample is a random variable because it is a function of the values of x_i which are random variables.

21. Probability Distribution - the probabilities associated with the values which a random variable can equal. If there are N values in a set, namely $X_1,\ldots,X_N$, and a random variable X can equal these values with probabilities $P_1,\ldots,P_N$, the probabilities constitute the probability distribution of X.

22. Estimator - a mathematical rule, method, or formula for estimating a population parameter from a sample. An estimate is a random variable when derived from a probability sample.

23. Ratio Estimator - an estimator involving ratios of random variables. Generally, a ratio estimator is one of two types: One is an average of ratios of two random variables such as $\frac{x_i}{y_i}$ for individual elements in a sample, and the other is the ratio of two estimates that are random variables such as $\frac{\bar{x}}{\bar{y}}$.

24. Sampling Distribution - the probability distribution of all possible estimates that an estimator might give under a specified sampling plan. To illustrate, reference is made to simple random sampling as defined above and to the sample average, $\bar{x}$, as an estimator of the population average, $\bar{X}$. For every possible sample of a given size that might occur, there is a value of $\bar{x}$. Each value of $\bar{x}$, has a probability of occurrence. These values of $\bar{x}$ and their probabilities of occurrence constitute the sampling distribution of $\bar{x}$ for simple random sampling.

It has been established on the basis of theory and empirical tests that the distribution of an estimate is approximately normal for a very large proportion of the estimates made from probability sample surveys. Most of the exceptions occur for estimates based on less than about 25 random selections. The distribution of an estimate (sampling distribution) is the center of attention of sampling specialists because it is information about this distribution that provides the basis for making statements concerning the accuracy or precision of an estimate, for determining sample sizes needed, and for determining the best of alternative sampling designs.

It is indeed fortunate, and a most remarkable fact, that one can make inferences about the sampling distribution for an estimator from a probability sample itself. It is also possible to infer from one sample what the sampling distribution would be like for samples of alternative sizes and design. Non-probability sampling does not furnish such capability.

25. Expected Value - (of an estimate) -- the average value of all possible estimates under specified conditions. The letter E is used to denote expected value. With reference to the notation used in the definition of "Probability Distribution," the expected value of a random variable X would be written: $E(X) = \Sigma P_i X_i$. Suppose p'_i is the probability that an estimate, x', is equal to x'_i where i is the index to all possible estimates that x' might be equal to. Mathematically, $E(x') = \Sigma p'_i x'_i$. That is, the expected value of an estimate, x', is the center (average) of its sampling distribution. Conceptually, if the sampling and estimation procedures are unbiased, the average of the sampling distribution is the result that would be obtained from a survey (census) of all sampling units in the frame using identical definitions, procedures, timing, field staff, adjustments for nonresponse etc., that were employed when the sample survey was conducted.

26. Random Error - (in an estimate) -- the difference between the estimate and its expected value. This difference is a random variable depending on the sample that happens to be selected. Let $e_i = x'_i - E(x'_i)$, where x'_i is an estimate and $E(x'_i)$ is the expected value of x'_i. Then, e_i is the random error in an estimate, x'_i, from a particular sample where i is an index of the samples that might have been selected.

27. Sampling Variance - (of an estimate) -- the variance of e_i described in the preceding definition. Mathematically, it is $E(e_i^2)$, the average square of the random errors, e_i. One goal in probability sampling is to be able to

obtain an estimate of the sampling variance, $E(e_i^2)$, from the sample. The term "variance of an estimate" is frequently used even though the variance among all possible estimates is referred to. Some writers prefer to use "variance of the estimator."

28. Standard Error - (of an estimate) -- the square root of the sampling variance. It pertains to the sampling distribution of an estimate, x', which is generally assumed to be normal. To illustrate, if x' is an estimate of the yield of corn per acre and the standard error of x' is 2 bushels, the probability is 2/3 that the estimate, x_i', will be within 2 bushels on either side of $E(x')$, the center of the sampling distribution. A more complete interpretation will be given later.

29. Relative Standard Error - (of an estimate) -- the standard error divided by the population parameter being estimated. It is usually expressed in percent. Relative standard error is estimated by dividing the estimate of the standard error by the estimate. "Standard error" is frequently used instead of "relative standard error." For example, "the standard error is 5 percent."

30. Coefficient of Variation - another name for "relative standard error."

31. Relative Variance - the square of the relative standard error. "Relvariance" is often used in lieu of "relative variance."

32. Measurement Error or Response Error- (for an individual element) -- the difference between an observed value, X_i, and a "true value," T_i. The measurement errors in statistics are functions of the differences $X_i - T_i$. The importance of measurement error varies widely with the nature of the characteristics, with the purpose of the survey, and with the magnitude of measurement error compared to other sources of error, particularly sampling error. Elementary sampling theory is developed under an assumption that the population set of values of X_i are fixed values.

33. Coverage Error - error caused by the omission or duplication of elements, or parts of elements, that are defined for inclusion in a probability sample survey or a census.

34. Nonresponse - refers to missing data. For various reasons, data might be missing for some elements. Methods of assigning values for missing data or methods of adjustment must be prescribed.

35. Bias - the difference between the expected value of an estimate and a parameter. If the average of the sampling distribution does not equal the

parameter, the estimate is biased. In practice there is no way of knowing whether the expected value of an estimate is exactly equal to the parameter, because the expected value and the parameter are unknown. Evidences of bias are manifest in various ways. Biases of major concern are possible biases resulting from response error, nonresponse, defects in the sampling frame and in the selection of a sample, the wording of questions, coverage errors and enumerator errors.

36. Unbiased Estimator - an estimator that is unbiased with regard to sampling and estimation methods. That is, with an unbiased estimator the center of the sampling distribution would be on target (equal to parameter) if the only source of error was sampling.

37. Unbiased Estimate - an estimate produced by an unbiased estimator. Such an estimate is "unbiased" only with regard to the techniques of selecting the sample and of estimation. It is not free of possible bias from other sources. The term has often been misleading.

38. Precision - variation of estimates obtained in repeated trials under the same conditions. Standard error is a measure of precision. Precision is an attribute of the sampling distribution of an estimator. It is the closeness of possible estimates under specified conditions to the center of the distribution.

39. Accuracy - total error is an estimate. It is the combination of errors from all resources including any biases as well as random components of error. An accurate estimate is a precise estimate that has little or no bias. It is closeness of an estimate to the parameter whereas "precision" is closeness to the expected value. "Accuracy of data" might refer to measurment error. "Accuracy" is a generic term.

40. Root Mean Square Error - a measure of accuracy. In terms of expected values of a random variable, it is

$$\sqrt{E(x' - T)^2} = \sqrt{\sigma^2_{x'} + b^2}$$

where x' is an estimate,

T is the parameter being estimated

$\sigma^2_{x'}$ is the sampling variance of x'

and $b = E(x') - T$, which is bias.

The expression $E(x' - T)^2$ represents the "mean square error" of an estimate, which is the average of the squares of deviations of all possible estimates from the parameter. Root mean square error, RMSE, is the square root of the

mean square error. The value of b can only be estimated, usually by special studies or comparisons of data from two or more sources.

41. Sampling Error - random error attributable to sampling. "Sampling error" is a general term. It is commonly used as a short term for longer expressions such as "sampling standard error" or "standard error of an estimate".

Error attributable to sampling can have variable and constant components. Likewise, error from sources other than sampling can also have variable and constant components. Generally, in practice neither variable components of errors nor biases are separable by sampling and nonsampling sources. The standard error of an estimate, as computed in practice, is a combination of variable components of error from all sources, sampling and nonsampling; and the bias (as defined above in 40) is a combination of all biases. The computed standard error is usually dominated by error associated with sampling; and, with carefully controlled probability sampling, the overall bias is likely to be comprised mostly of biases associated with nonsampling sources.

When considered in detail and exactly in a practical setting, error concepts become complex and difficult to communicate. In the interest of simplicity and owing to the problem of communicating technical concepts, standard errors of estimates are usually described in statistical reports simply as measures of "sampling error", even though variable error from sources other than sampling are embraced in the computed estimates of standard errors. Additional statements are often made pointing out that nonsampling error also exists and is not included in the standard error of the estimate. In this context, "nonsampling error" refers to any biases that might exist.

42. Nonsampling Error - a general term applying to all sources of error other than error attributable to sampling. It includes error from any defects in the sampling frame, response error, and mistakes in processing the data.

43. Sampling Efficiency - the variance of estimates for one sampling plan compared to another assuming equal sample sizes or equal costs.

44. Domains - a general term for subdivisions of a population for which estimates are made. Domain specifications have an important bearing on sample design and estimation procedures.

45. Self-Weighted Sample - a sample wherein all elements of the population have an equal chance of being in the sample, so the sample data need not be weighted. Reasons for designing a sample that is not self-weighted are usually either matters of sampling efficiency or to better serve the objective of domain estimates.

46. Cluster Sampling - the selection of clusters of elements. That is, the sampling units are "clusters" of elements. Identification of elements belonging to each sampling unit (cluster) in a sample is often a part of the field operations. There must be rules for associating each element of the population with one and only one cluster.

47. Area Sampling - a special case of cluster sampling. The sampling units are areas of land, commonly called segments, which have identifiable boundaries. The idea is to divide the entire land area of the population to be surveyed into segments and to select a sample of such segments. The process of area sampling is usually accomplished by selecting the sample in stages which avoids the necessity of dividing the entire population into segments.

48. Intraclass Correlation Coefficient - a measure of the degree to which elements within clusters (sampling units) are alike. When cluster sampling is used, the sampling variance depends on the intraclass correlation. High intraclass correlation means low sampling efficiency (high sampling variance) compared to a sample with individual elements as the sampling units, assuming the same number of elements in both samples.

49. Stratification - the classification of sampling units in a population into groups called strata. The strata might be defined as domains for which estimates are derived, or as homogeneous groups of sampling units to reduce sampling error, or for both purposes. Stratification is used in most sampling plans.

50. Stratified Random Sampling - selecting a probability sample from each stratum of a stratified population. Sampling variance depends on the variation among sampling units within strata. Therefore, one goal in stratification is to form strata so the variation among sampling units within strata is small. Sampling variance is also effected by how the sample is allocated to the strata. Gains in sampling efficiency from stratification are generally moderate to substantial and rarely negligible or negative.

51. Post Stratification - a technique of estimation. Stratification takes place after the data are collected. The sampling units or elements in the sample are classified into "strata" and the sample is expanded stratum by stratum provided the population numbers of sampling units or elements are known for the strata.

52. Collapsed Strata - groups of two or more strata. When only one random selection is made from a stratum, approximate methods of estimating sampling error must be used. Similar strata are combined to form "collapsed strata." Variation among sampling units within "collapsed strata" is the basis for estimating sampling error. The number of original strata in a collapsed stratum is usually limited to two. In general, this method will overestimate sampling errors by a small amount.

53. Systematic Sample - a sample obtained by selecting every nth element in an array of elements of the population. A random starting point between 1 and n is selected. According to the definition of a sampling unit (Def. in 9), a systematic sample would consist of only one sampling unit which is a group of elements equally distributed through the file. The sampling variance could be much smaller or larger than the sampling variance for simple random sampling. Unbiased estimation of sampling variance is impossible when there is only one random start.

54. Sampling With PPS - (Probability Proportional to Size) -- a method of sampling wherein a sampling unit has a probability of selection that is proportional to its "size". This technique is generally used in multi-stage sampling. It can also be a valuable technique in single-stage sampling. Sampling variance is quite sensitive to how well the measure of size fills conditions for sampling with PPS to be effective.

55. Multi-Stage Sampling - two or more stages of sampling. To illustrate, there are many examples of multi-stage sampling in surveys to estimate crop yields. At the first stage of sampling, an area sample could be selected. All fields in each area sampling unit would be identified and the crop and acreage of each field recorded. For purposes of estimating crop acreage, the sample is single stage. But, to estimate yields, a sample of fields might be selected from the area sample and within the sample fields a sample of plots could be selected in which various measurements are taken for estimating yield. In this case, there are three stages of sampling: The first-stage sampling units are area segments, the second-stage units are fields, and the third-stage units are plots. There is a component of sampling error associated with each stage of selection.

56. Primary Sampling Units - sampling units at the first stage of sampling in a multi-stage sampling plan.

57. Ultimate Cluster - the sample from a primary sampling unit. The term was coined owing to a need in multi-stage sampling to make frequent reference to the sample from a primary sampling unit. In multi-stage sampling, standard errors of estimates may be estimated from the variation among ultimate clusters within strata used at the first stage of sampling.

58. Subsampling - a general term for selecting a sample from a sample. "Subsampling" has often been used instead of "two-stage sampling", especially when the primary sampling units are small.

59. Double Sampling - sampling in two phases: Phase one, the collection of a small amount of data from all units in a sample; and, Phase two, the collection of more detailed data from a subsample. The purpose of the first phase is to improve the sample for the second phase and achieve more accurate estimates for characteristics included in the second phase than would be possible from a single-phase sample at the same cost.

60. Replicated Sampling - the selection of more than one sample under the same sampling plan. Each sample provides estimates of the parameters. When replicated samples are selected independently, a very simple method is provided for estimating the standard error of an estimate.

61. Multi-Frame Sampling - joint use of more than one sampling frame. Collectively, the frames should include all elements of the population. Some elements will be included in two or more frames. Probably, the most common application is two frame sampling, where one frame is an area sampling frame that includes all elements and the second frame is an incomplete list of elements that are most important for the survey.

Alphabetic Listing of Terms

Accuracy	(39)	Ratio Estimator	(23)
Area Sampling	(47)	Relative Standard Error	(29)
Bias	(35)	Relative Variance	(31)
Census	(12)	Replicated Sampling	(60)
Characteristic	(4)	Response Error	(32)
Cluster Sampling	(46)	Root Mean Square Error	(40)
Coefficient of Variation	(30)	Sample Design	(11)
Collapsed Strata	(52)	Sampling Distribution	(24)
Coverage Error	(33)	Sampling Efficiency	(43)
Domains	(44)	Sampling Error	(41)
Double Sampling	(59)	Sampling Frame	(14)
Estimator	(22)	Sampling Unit	(9)
Expected Value	(25)	Sampling Variance	(27)
Frame Units	(15)	Sampling With PPS	(54)
Intraclass Correlation Coefficient	(48)	Self-Weighted Sample	(45)
Measurement Error	(32)	Simple Random Sampling	(10)
Multi-Frame Sampling	(61)	Standard Error	(28)
Multi-Stage Sampling	(55)	Statistical Population	(7)
Non-Response	(34)	Statistical Survey	(1)
Non-Sampling Error	(42)	Statistics	(13)
Population	(2)	Stratification	(49)
Population Parameters	(18)	Stratified Sampling	(50)
Post Stratification	(51)	Subsampling	(58)
Precision	(38)	Systematic Sample	(53)
Primary Sampling Units	(56)	True Values	(19)
Probability Distribution	(21)	Ultimate Cluster	(57)
Probability of Inclusion	(16)	Unbiased Estimate	(37)
Probability of Selection	(17)	Unbiased Estimator	(36)
Probability Sample	(8)	Unit of Analysis	(6)
Random Error	(26)	Unit of Observation	(5)
Random Variable	(20)	Universe	(3)