

Appendix F

Definitions

Area sampling – a method of sampling a geographic area that is subdivided into smaller areas, such as counties, townships, or city blocks. A random sample of areas is selected and either surveyed 100 per cent, or further sampled.

Attribute – a qualitative characteristic of interest associated with sampling units.

Attribute sampling – a type of sampling that answers the questions “how many?” or “what percentage?”

Audited Amount – in financial auditing, is the amount that should be in the taxpayer’s records.

Audit risk – is the risk that there is tax misstatement and that the auditor will not detect that error.

Bias – is not the same as sampling error (precision) – these two concepts should not be confused. It has a variety of meanings, but generally describes something that should be avoided or minimized, if possible. Following are some uses of the term in statistical sampling:

- With regard to estimation, an estimator is said to be biased if the expected value is not the same as the parameter it is intended to estimate.
- Sampling is biased when the chances of selection are unknown, or where some units are unintentionally favored over other units with unplanned probability. Note that in some sampling procedures, such as stratified sampling or sampling using probability proportional to size, the units in the sampling frame will likely have different but known chances of selection. Therefore if these procedures are preformed correctly and according to plan, they will not be biased with regard to sampling.
- Another form of sampling bias occurs when the sampling frame does not adequately correspond to the target population.
- A biased procedure is defined as “the difference between the results that it produces and the results that would be produced had a preferred procedure been used instead”. An example of this is where the auditor fails to correctly value the sample units.

- Another example of a biased procedure is extending the results of an evaluation of one population that was sampled, to another population not sampled.

Block sample – a non-probabilistic sample based on a portion of the population, common in audit sampling, where a large segment of the population, based on convenience, is selected. Source records for block samples are generally located in one place or are physically continuous in nature. A cluster sample is similar to a block sample in that the sample elements are also continuous in some manner. However, a cluster sample, as defined herein, involves selection using probabilistic methods and where there are generally a number of clusters drawn from the population (block samples will often have only a few segments, or only one segment, of the population). A block sample, as defined herein and as generally applied in audit sampling, is a judgmental sample.

Bootstrapping – statistical inference based on the results of repeatedly resampling the sample. This procedure can be used to establish a confidence interval from a probability sample without any assumptions regarding the sampling distribution.

Cluster sample – is a population divided into groups, or clusters. A simple random sample is taken, where the cluster is the sampling unit. Those clusters chosen in the sample are then audited in their entirety.

Coefficient of variation (CV) – is the ratio of the standard deviation to the corresponding mean. The population CV measures the relative dispersion of the population distribution. The CV of an estimate is the standard error divided by the estimate.

Confidence interval – is the range of values between the lower confidence limit (LCL) and the upper confidence level (UCL) derived from the sample, which contains the true population value with a specified confidence level.

Confidence level – represents the frequency that confidence intervals calculated from all possible samples contain the true population amount.

Correlation – is a measure of the degree in which two quantities are linearly related. This relationship could be one of dependence or association. Correlation in a population can be measured by the correlation coefficient, ρ (pronounced “row”). The extreme values for ρ of -1 or +1 signify an exact linear relationship. If $\rho = 0$, then the two quantities have no linear

relationship. The sample correlation is $\hat{\rho}$, which also ranges from -1 to 1. A formula for $\hat{\rho}$ is included in Appendix A.

Correspondence – the method of matching random numbers to sampling units from the sampling frame.

Decision rules – special valuation rules employed by the auditor in valuing the sample whereby taxable error has only one chance of being anything other than zero, avoiding selection bias and possible impairment of the sample results.

Detail – a review of all transactions within an audit by the auditor using thorough, complete, and consistent procedures throughout the examination. Statisticians often refer to a detail as a census.

Difference estimator – is an evaluation procedure in which the average taxable error (or other unknown) from the sample is used to estimate the average taxable error (or other unknown) for the population. The total taxable error for the population is calculated by multiplying the average taxable error from the sample by the population count.

Discovery sampling – a type of sampling that attempts to determine whether a specific condition exists in a population.

Distribution – refers to the scattering or diffusion of values in a population.

Error rate – the occurrence rate of sampling units in the population (or sample) that have a non-zero taxable error value.

Estimator – is some mathematical expression used to estimate a parameter from a population.

Exponentially distributed – an asymmetrically distributed distribution, where the values increase or decrease at a geometric rate.

Finite population correction (fpc) – is an adjustment to the standard error of the sample mean when sampling is without replacement from a finite population.

Frequency distribution – is a tabular representation of a population where the population is divided into classes or ranges and the number of population units falling into each class is counted and shown in the table.

Heterogeneous – exhibiting characteristics of a different or unlike nature.

Homogeneous – uniform or consistent with respect to a particular characteristic.

Hypergeometric distribution – The exact sampling distribution for the sample occurrence rate when an unrestricted random sample (SRS) of a given sample size and population is selected. The distribution will be different for any change in the sample size or population size.

Judgment sample – a sample picked from the population by the subjective decision of an individual where the chance of selection is not known.

Kurtosis – the relative peakedness or flatness of a distribution.

Lower confidence limit – (LCL) is the point estimate less the precision amount and is the smallest value in the range of values, the confidence interval, which contains the true population value with a specified confidence level.

Mean-per-unit estimator – an estimator that backs into the total taxable error value of the population by first estimating the total audited value for the population. The total taxable error value equals the total invoice value less the total audited value. The total audited value is estimated by taking the average audited value from the sample multiplied by the population count. Each individual audit value for the sample is computed by subtracting the taxable error value for that item with the invoice value for that item.

Mixture – a distribution that appears to be a mixture of two or more common types.

Multi-phase sampling – see multistage sample.

Multistage sample – similar to a cluster sample, but instead of auditing the selected clusters in their entirety (the clusters are then referred to as primary or first-stage units), the selected clusters are sub-sampled. The audit is performed on the sub-sampled units (called secondary or second-stage units).

Non-sampling error – an error that is encountered whether the population is sampled or not.

Normal distribution – (also known as the Gaussian distribution) is an important distribution in statistical theory used to estimate probabilities. It is symmetric and bell-shaped distribution and is the approximate sampling distribution for many statistical estimators.

Occurrence rate – the rate of occurrence in the population (or sample) of sample units exhibiting an attribute.

Parameter – a numerical quantity that describes the population. Parameters are often denoted with a Greek letter.

Period – is the maximum length of the sequence numbers before a PRNG begins to repeat itself.

Point estimate – is an estimate of a parameter of a population. Most often in tax auditing, the point estimate of interest is an estimate of total error.

Population – all transactions or records in an audit examination. Sometimes the population is referred to as the universe.

Population mean – or the true mean value of the population.

Population standard deviation – a measure of variation of the population, on a scale the same as the population mean, and is the square root of the population variance.

Population variance – or true variance, is a measure of variation of all the values in a population in relationship to the population mean.

Precision – see the definition of “precision amount”. The term is also known as sampling error.

Precision amount – or just “precision” and sometimes referred to as “sampling error” is the measure of how close a sample estimate is from the corresponding population characteristic. It is computed by multiplying the standard error of the estimate by a factor determined by the desired confidence.

Probability proportional to size sample – is a type of sample where the probability of selection is proportional to the size (or dollar value) of the unit. Some references call this dollar unit sampling (DUS) or monetary unit sampling (MUS). Mechanically, in PPS designs, each dollar in the population has an equal chance of selection. The randomization or matching to the random numbers is tied to each dollar in the population. If a particular dollar is selected into the sample, the entire document is pulled into the sample.

Probability sample – a sample where the chance of selection of every item in the population has a known, but not necessarily equal chance of selection (contrast this definition with that of a judgmental sample below).

Pseudorandom number generator (PRNG) – is an algorithm used to generate a sequence of numbers that approximate the properties of random numbers. The numbers supplied are a result

of a deterministic process (therefore can not be truly random as they are predictable), but are said to have qualities of randomness, and are therefore referred to as pseudorandom.

Random numbers – a set of numbers that can occur uniformly over a given range (such as between the range of 0 and 1) but have no predictable or traceable pattern.

Random sample – (often “simple random sample”) when taken from a finite population, each possible sample for sample size n from a population of N is possible and has an equal chance of being selected.

Range – the difference between the highest and lowest values in a group of items.

Ratio estimator – the population ratio, R , equals the total taxable error divided by the total invoice amount. The sample ratio, \hat{R} , is used to estimate the population ratio. The sample ratio is derived by dividing the total taxable error in the sample by the total invoice amount in the sample. The sample ratio is used to estimate total taxable error by taking \hat{R} and multiplying it by the known total population invoice value.

Regression estimator – uses the linear relationships of the taxable error values and the invoice values in the sample, in addition to the known total population invoice value, to estimate the total taxable error in the population.

Relative precision – (sometimes precision percentage) is the precision amount expressed in relative terms to the point estimate.

RHC sampling – is a multistage form of sampling, where the primary units (first stage) are selected using probabilistic methods that consider the size of the primary unit. If the sample has three stages, it can be used in both the primary and secondary stages. RHC sampling allows sampling without replacement, but has some of the benefits of PPS sampling (which requires sampling with replacement).

Sample - a part from a larger whole or group selected for inspection.

Sampling - the act, process or technique of selecting a sample.

Sample design – a plan for sampling a population specified before sampling commences, often referred to as a ‘sampling plan’ (a sample plan includes more information, such as sample size). Various sample designs include cluster sampling, multi-stage sampling, simple random

sampling, and stratified random sampling. Key elements of the sample plan include sample design as well as identification of the sampling frame and sampling unit, sample size, determination of the source of random numbers, definition of the characteristic being estimated, and identification of the estimator used to project the sample results.

Sampling error – another term for precision, and is the difference between a value from a population, usually not known, and an estimate using a sample from that population.

Sampling frame (or frame) – is the list or file sampled and the means by which the target is sampled.

Sampling distribution – is the distribution of all averages, totals, percentages, or other statistics for all possible samples of a given sample size for a certain population. If the sample is “large” then the sampling distribution will approximate the normal distribution.

Sampling risk – is the chance that the confidence interval will not contain the population parameter of interest.

Sampling unit – is each individual element of the sampling frame that can be selected into the sample.

Sampling with replacement – a sampling procedure where individual sampling units are returned to the population before selecting subsequent units.

Sampling without replacement – a sampling procedure where individual sampling units are not returned to the population before selecting subsequent units.

Seed – is an arbitrary starting point to initiate an algorithm that supplies pseudorandom numbers.

Sequential sample – a form of sampling where the sample is drawn one at a time, and after each observation, a decision is made whether to continue to sample or accept the results of the sample. This necessarily means that the sample was performed in the random order of the sample units.

Simple random sample (SRS or unrestricted random sample) – is a sample drawn from a population so that each sample of the same size has the same probability of being selected. An unrestricted random is presumed to be taken without replacement unless stated otherwise.

Skew – an unsymmetrical frequency distribution where percentage of values above the mean differs from the percentage of values below the mean.

Standard deviation – is a measure of variability within a population or a sample. The standard deviation is the square root of the variance.

Standard error – is the standard deviation of a sampling distribution and measures the variability of the estimates.

Statistic – is a numerical quantity that describes or comes from a sample.

Statistical inference – any form of reasoning that estimates population parameters from sample data, including any generalization, prediction, estimate or decision based on a sample.

Statistical sampling - means any approach to sampling that has the following characteristics:

- (1) Use of a probability sample; and
- (2) Use of probability theory to evaluate sample results, including measurement of sampling risk.

A sampling approach that does not have characteristics (1) and (2) is considered non-statistical sampling.

Stratified random sample – sampling a population divided up into different non-overlapping groups, or strata. A random sample is then taken from each stratum.

Stratifying – dividing a population, or sampling frame, into non-overlapping groups, or strata. In most cases, stratifying is done to take more than one sample – although in some cases the auditor may stratify and take only one sample from one of the groups (considered simple random sampling).

Tail – in a skew distribution, this describes one side of the distribution. Positive skewness or right skew distributions show values extending out away from the mean farther toward the right and are said to have a “right tail”. Here, values above the mean occur less frequently than values below the mean. Negative skewness, or left skew distributions show the values moving out farther away to the left, and are said to have a “left tail” where values above the mean occur more frequently than values below the mean..

Target population – is the population that is of audit interest.

Upper confidence limit – is the point estimate plus the precision amount.

Variable sampling – a type of sampling that reaches conclusions on monetary values or other amounts from a population.

Variance – a measure of variability within a population or a sample. The variance is the standard deviation squared.

Variation – measurement of how spread out or dispersed the values are in a population or sample.