Designation: E1402 – 13 (Reapproved 2018)

An American National Standard

Standard Guide for Sampling Design[1]

This standard is issued under the fixed designation E1402; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This guide defines terms and introduces basic methods for probability sampling of discrete populations, areas, and bulk materials. It provides an overview of common probability sampling methods employed by users of ASTM standards.

1.2 Sampling may be done for the purpose of estimation, of comparison between parts of a sampled population, or for acceptance of lots. Sampling is also used for the purpose of auditing information obtained from complete enumeration of the population.

1.3 No system of units is specified in this standard.

1.4 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety, health, and environmental practices and determine the applicability of regulatory limitations prior to use.*

1.5 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

D7430 Practice for Mechanical Sampling of Coal

E105 Practice for Probability Sampling of Materials

E122 Practice for Calculating Sample Size to Estimate, With Specified Precision, the Average for a Characteristic of a Lot or Process

E141 Practice for Acceptance of Evidence Based on the Results of Probability Sampling

E456 Terminology Relating to Quality and Statistics

3. Terminology

3.1 *Definitions*—For a more extensive list of statistical terms, refer to Terminology E456.

3.1.1 *area sampling, n*—probability sampling in which a map, rather than a tabulation of sampling units, serves as the sampling frame.

3.1.1.1 *Discussion*—Area sampling units are segments of land area and are listed by addresses on the frame prior to their actual delineation on the ground so that only the randomly selected ones need to be exactly identified.

3.1.2 *bulk sampling, n*—sampling to prepare a portion of a mass of material that is representative of the whole.

3.1.3 *cluster sampling, n*—sampling in which the sampling unit consists of a group of subunits, all of which are measured for sampled clusters.

3.1.4 *frame, n*—a list, compiled for sampling purposes, which designates all of the sampling units (items or groups) of a population or universe to be considered in a specific study.

3.1.5 *multi-stage sampling, n*—sampling in which the sample is selected by stages, the sampling units at each stage being selected from subunits of the larger sampling units chosen at the previous stage.

3.1.5.1 *Discussion*—The sampling unit for the first stage is the primary sampling unit. In multi-stage sampling, this unit is further subdivided. The second stage unit is called the secondary sampling unit. A third stage unit is called a tertiary sampling unit. The final sample is the set of all last stage sampling units that are obtained. As an example of sampling a lot of packaged product, the cartons of a lot could be the primary units, packages within the carton could be secondary units, and items within the packages could be the third-stage units.

3.1.6 *nested sampling, n*—same as *multi-stage sampling*.

3.1.7 *primary sampling unit, PSU, n*—the item, element, increment, segment or cluster selected at the first stage of the selection procedure from a population or universe.

3.1.8 *probability proportional to size sampling, PPS, n*—probability sampling in which the probabilities of selection of sampling units are proportional, or nearly proportional, to a quantity (the "size") that is known for all sampling units.

3.1.9 *probability sample, n*—a sample in which the sampling units are selected by a chance process such that a

[1] This guide is under the jurisdiction of ASTM Committee E11 on Quality and Statistics and is the direct responsibility of Subcommittee E11.10 on Sampling / Statistics.

Current edition approved Nov. 1, 2018. Published November 2018. Originally approved in 2008. Last previous edition approved in 2013 as E1402 – 13. DOI: 10.1520/E1402-13R18.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

specified probability of selection can be attached to each possible sample that can be selected.

3.1.10 *proportional sampling, n*—a method of selection in stratified sampling such that the proportions of the sampling units (usually, PSUs) selected for the sample from each stratum are equal.

3.1.11 *quota sampling, n*—a method of selection similar to stratified sampling in which the numbers of units to be selected from each stratum is specified and the selection is done by trained enumerators but is not a probability sample.

3.1.12 *sampling fraction, f, n*—the ratio of the number of sampling units selected for the sample to the number of sampling units available.

3.1.13 *sampling unit, n*—an item, group of items, or segment of material that can be selected as part of a probability sampling plan.

3.1.13.1 *Discussion*—The full collection of sampling units listed on a frame serves to describe the sampled population of a probability sampling plan.

3.1.14 *sampling with replacement, n*—probability sampling in which a selected unit is replaced after any step in selection so that this sampling unit is available for selection again at the next step of selection, or at any other succeeding step of the sample selection procedure.

3.1.15 *sampling without replacement, n*—probability sampling in which a selected sampling unit is set aside and cannot be selected at a later step of selection.

3.1.15.1 *Discussion*—Most samplings, including simple random sampling and stratified random sampling, are conducted by sampling without replacement.

3.1.16 *simple random sample, n*—(without replacement) probability sample of n sampling units from a population of N units selected in such a way that each of the $\frac{N!}{n!(N-n)!}$ subsets of n units is equally probable – (with replacement) a probability sample of n sampling units from a population of N units selected in such a way that, in order of selection, each of the N^n ordered sequences of units from the population is equally probable.

3.1.17 *stratified sampling, n*—sampling in which the population to be sampled is first divided into mutually exclusive subsets or strata, and independent samples taken within each stratum.

3.1.18 *systematic sampling, n*—a sampling procedure in which evenly spaced sampling units are selected.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *address, n*—(sampling) a unique label or instructions attached to a sampling unit by which it can be located and measured.

3.2.2 *area segment, n*—(area sampling) final sampling unit for area sampling, the delimited area from which a characteristic can be measured.

3.2.3 *composite sample, n*—(bulk sampling) sample prepared by aggregating increments of sampled material.

3.2.4 *increment, n*—(bulk sampling) individual portion of material collected by a single operation of a sampling device.

3.3 *Symbols:*

N = number of units in the population to be sampled.
n = number of units in the sample.
Y_i = quantity value for the *i-th* unit in the population.
y_i = quantity observed for *i-th* sampling unit.
$\bar{Y}$ = average quantity for the population.
$\bar{y}$ = average of the observations in the sample.
X_i = value of an auxiliary variable for the *i-th* unit in the population.
x_i = value of an auxiliary variable for the *i-th* sampling unit.
P = population proportion of units having an attribute of interest.
p = sample proportion.
f = sampling fraction.
s = sample standard deviation of the observations in the sample.
s^2 = sample variance of the observations in the sample.
$SE(\bar{y})$ = standard error of an estimated mean $\bar{y}$.

4. Significance and Use

4.1 This guide describes the principal types of sampling designs and provides formulas for estimating population means and standard errors of the estimates. Practice E105 provides principles for designing probability sampling plans in relation to the objectives of study, costs, and practical constraints. Practice E122 aids in specifying the required sample size. Practice E141 describes conditions to ensure validity of the results of sampling. Further description of the designs and formulas in this guide, and beyond it, can be found in textbooks (**1-10**).[3]

4.2 Sampling, both discrete and bulk, is a clerical and physical operation. It generally involves training enumerators and technicians to use maps, directories and stop watches so as to locate designated sampling units. Once a sampling unit is located at its address, discrete sampling and area sampling enumeration proceeds to a measurement. For bulk sampling, material is extracted into a composite.

4.3 A sampling plan consists of instructions telling how to list addresses and how to select the addresses to be measured or extracted. A frame is a listing of addresses each of which is indexed by a single integer or by an n-tuple (several integer) number. The sampled population consists of all addresses in the frame that can actually be selected and measured. It is sometimes different from a targeted population that the user would have preferred to be covered.

4.4 A selection scheme designates which indexes constitute the sample. If certified random numbers completely control the selection scheme the sample is called a probability sample. Certified random numbers are those generated either from a table (for example, Ref (**11**)) that has been tested for equal digit frequencies and for serial independence, from a computer

[3] The boldface numbers in parentheses refer to a list of references at the end of this standard.

program that was checked to have a long cycle length, or from a random physical method such as tossing of a coin or a casino-quality spinner.

4.5 The objective of sampling is often to estimate the mean of the population for some variable of interest by the corresponding sample mean. By adopting probability sampling, selection bias can be essentially eliminated, so the primary goal of sample design in discrete sampling becomes reducing sampling variance.

5. Simple Random Sampling (SRS) of a Finite Population

5.1 Sampling is without replacement. The selection scheme must allocate equal chance to every combination of n indexes from the *N* on the frame.

5.1.1 Make successive equal-probability draws from the integers 1 to *N* and discard duplicates until n distinct indexes have been selected.

5.1.2 If the *N* indexed addresses or labels are in a computer file, generate a random number for each index and sort the file by those numbers. The first n items in the sorted file constitute a simple random sample (SRS) of size *n* from the *N*.

5.1.3 A method that requires only one pass through the population is used, for example, to sample a production process. For each item, generate a random number in the range 0 to 1 and select the *i-th* item when the random number is less than $(n - a_i)/(N - i + 1)$, where a_i is the number of selections already made up to the *i-th* item. For example, the first item ($i = 1$ and $a_1 = 0$) is selected with probability *n/N*.

5.2 The quantities observed on the variable of interest at the selected sampling units will be denoted $y_1, y_2, \ldots, y_n$. The estimate of the mean of the sampled population is

$$\bar{y} = \sum y_i / n \quad (1)$$

The standard error of the mean of a finite population using simple random sampling without replacement is:

$$SE(\bar{y}) = s\sqrt{(1 - f)/n} \quad (2)$$

where $f = n/N$ is the sampling fraction and s^2 is the sample variance (*s*, its square root, is sample standard deviation).

$$s^2 = \sum (y_i - \bar{y})^2/(n - 1) \quad (3)$$

The population mean that $\bar{y}$ estimates is:

$$\bar{Y} = \sum_{i=1}^{N} Y_i / N \quad (4)$$

The expected value of s^2 is the finite population variance defined as:

$$S^2 = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2/(N - 1) \quad (5)$$

5.3 *Finite Population Correction*—The factor $(1 - f)$ in Eq 2 is the finite population correction. In conventional statistical theory, the standard error of the average of independent, identically distributed random variables does not include this factor. Conventional statistical theory applies for random sampling with replacement. In sampling without replacement from a finite population, the observations are not independent. The finite population correction factor depends on (*a*) the population of interest being finite, (*b*) sampling being without errors and measurements for any sampled item being assumed completely well defined for that item. When the purpose of sampling is to understand differences between parts of a population (analytic as opposed to enumerative, as described by Deming (**4**)), actual population values are viewed as themselves sampled from a parent random process and the finite population correction should not be used in making such comparisons.

5.4 *Sample Size*—The sample size required for a sampling study depends on the variability of the population and the required precision of the estimate. Refer to Practice E122 for further detail on determining sample size. Eq 2 can be developed to find required sample size. First, the user must have a reasonable prior estimate s_0 of the population standard deviation, either from previous experience or a pilot study. Solving for *n* in Eq 2, where now $SE(\bar{y})$ is the required standard error, gives:

$$n = \frac{n_o}{1 + n_o/N} \quad \text{where:} n_o = s_o^2/SE(\bar{y})^2 \quad (6)$$

5.5 *Estimating a Proportion*—Formulas 1 through 5 serve for proportions as well as means. For an indicator variable Y_i which equals 1 if the *i-th* unit has the attribute and 0 if not, the population proportion $P = \bar{Y}$ can be recognized as the average of ones and zeros. The sample estimate is the sample proportion $p = \bar{y}$ and the sample variance is $s^2 = np(1 - p)/(n - 1)$.

5.6 *Ratio Estimates*—An auxiliary variable may be used to improve the estimate from an SRS. Values of this variable for each item on the frame will be denoted X_i. Specific knowledge of each and every X_i is not necessary for ratio estimation but knowing the population average $\bar{X}$ is. The observed values x_i are needed along with the y_i, where the index *i* goes from $i = 1$ to $i = n$, the sample size. The estimated ratio is $\hat{R} = \bar{y}/\bar{x}$ and the improved ratio estimate of $\bar{Y}$ is $\bar{X}\bar{y}/\bar{x}$. The estimated standard error of the ratio estimate of $\bar{Y}$ is:

$$SE\left(\bar{X}\hat{R}\right) = \sqrt{\frac{1 - f}{n}\sum\left(y_i - \hat{R}x_i\right)^2/(n - 1)} \quad (7)$$

5.6.1 The ratio estimator works best when the relation of *X*-values to *Y*-values is approximately linear through the origin with the variance of *Y* for given *X* approximately proportional to *X*. Other estimates using the auxiliary variable include regression estimators and difference estimators (**2**). The best form of estimate depends on the relation of *X* to *Y* values and the relation between the variance of *Y* for given *X*.

6. Systematic Selection (SYS)

6.1 For systematic selection of a sample of n from a list of *N* sampling units when $N/n = k$ is integer, a random integer between 1 and *k* should be selected for the start and every *k-th* unit thereafter. When *N/n* is not integer, then a random integer between 1 and *N* should be selected for the start and the nearest integer to *N/n* added successively, subtracting *N* when exceeded, to get selected units. Multiple starts should be used to create replicated samples (Practice E141) for estimating sampling error if sample size *n* is large.

6.2 If an auxiliary variable, the X_i of 5.6, is available, it can be used to sort the units of the frame so that a systematic sample will contain a balanced cross section of the X_i values.

6.3 The sample average $\bar{y}$ is an unbiased estimate of the population mean. An estimate of the standard error of $\bar{y}$ based on the first differences is:

$$SE(\bar{y}) = \sqrt{\frac{1}{2n}\sum_{j=2}^{n}(y_j - y_{j-1})^2/(n-1)} \tag{8}$$

6.4 When K replicated subsamples are used, each subsample mean, $\bar{y}_k$, estimates the population mean and the average of all, $\bar{\bar{y}}$, is the overall estimate. A preferred number of replicate subsamples is five to ten. The standard error is:

$$SE(\bar{\bar{y}}) = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(\bar{y}_k - \bar{\bar{y}})^2/(K-1)} \tag{9}$$

7. Probability Proportional to Size (PPS) Sampling

7.1 When the frame lists an auxiliary ("size") variable X_i for every address and the X-values are correlated with the Y-values, then it may be efficient to select the sampling units with probability proportional to the X_i values.

7.2 Cumulate sizes X_i to get $C_i = \Sigma X_j$ summing over j less than or equal to i. If the X_i are decimal, multiply by a power of ten to make usable integers. C_N is the overall sum. A random integer, say r, in the range 1 to C_N will lie in some interval $C_{i-1} < r <= C_i$ and selects unit i with probability proportional to X_i. Generating n such integers with replacement selects a PPS with replacement sample. Duplicated selections, if any, are measured again.

7.3 Data from a with-replacement PPS sample are converted to ratios $z_i = y_i/x_i$, which are independently and identically distributed with mean equal to the sum of Y-values divided by the sum of X-values. The estimate of the population mean, $\bar{Y}$, is:

$$\bar{y}_{PPS} = \bar{z}\bar{X} \tag{10}$$

with standard error:

$$SE(\bar{y}_{PPS}) = \bar{X}\sqrt{1/n\sum_{i=1}^{n}(z_j - \bar{z})^2/(n-1)} \tag{11}$$

NOTE 1—Simple PPS sampling without replacement can be conducted by independent draws selecting sampling unit i, if it remains unselected, at each step with probability proportional to X_i. However, the resulting probabilities of inclusion in the sample for each item are not exactly proportional to their size. Modified PPS schemes are reviewed by Brewer and Hanif (**12**).

7.4 A PPS sampling without replacement method with the property that inclusion probabilities are proportional to sizes can be accomplished. Form cumulative sums C_i following 7.2. If there are large units with size $X_i > C_N / n$ then they must be selected for sure, removed from the probability sampling frame, and cumulative sums recomputed to select the remainder of the sample. Systematically sample n integers from the cumulative size range 1 to C_N in accord with 6.1 and then measure the units thus selected.

7.4.1 The estimate of the population mean for this systematic PPS without replacement sampling is:

$$\bar{y}_{PPS} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}\right)\bar{X} \tag{12}$$

where $\bar{X} = C_N/N$ is the population mean size. $\pi_i = \frac{nx_i}{C_N}$ is the inclusion probability for unit i.

The first formula of Eq 12 is known as the Horvitz-Thompson estimate (**13**). An approximate formula for the standard error of $\bar{y}_{PPS}$ is due to Hartley and Rao (**14**). If selection probabilities are exactly proportional to Y_i, then the standard error of the PPS estimate $\bar{y}_{PPS}$ is zero.

$$SE(\bar{y}_{PPS}) = \sqrt{\frac{1}{N^2(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left[1-(\pi_i+\pi_j)+\sum_{k=1}^{N}\pi^2{}_k/n\right]\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2} \tag{13}$$

7.5 An alternative to this form of unequal probability sampling is to stratify the population by size, and conduct stratified sampling with the size categories as strata.

8. Stratified Sampling

8.1 The frame for stratified sampling includes division of the sampling units into disjoint and exhaustive subsets of similar sampling units, called strata. Addresses are two-digit indexes where the first number refers to the stratum while the second identifies the sampling unit within each stratum. Stratified sampling requires that some item be sampled from every stratum on the stratified frame.

8.2 After listing the sampling units in each stratum on a frame, the selection is made of n_1 from the N_1 in the first stratum, of N_2 from N_2 in the second, and so on to n_L from N_L in the last stratum.

8.3 The numbers n_1, n_2, …, n_L are called an allocation. Common allocations are:

(1) Proportional to N_h,

(2) Neyman (**15**), proportional to $N_h S_h$ (where S_h is stratum standard deviation),

(3) Optimum, proportional to $N_h S_h/\sqrt{C_h}$ where C_h is cost per observation in stratum h,

(4) Equal, all n_h equal, and

(5) Compromise, proportional to $N_h^{0.5}$ (exponents other than 0.5 can also be used).

8.4 The first three require increasing amounts of preliminary information so that the second and third are seldom used. Proportional allocation has the convenient property that the estimate of the overall population mean is the unweighted sample average. Equal allocation is appropriate if comparisons between strata or means for individual strata are of interest (Practice E105). The compromise allocation mediates between goals of estimating stratum averages and estimating the overall population mean. Values of the exponent less than 0.5 better estimate stratum mean differences. Exponent 0.0 gives equal allocation. Values greater than 0.5 are better for estimating the overall mean. Exponent 1.0 gives proportional allocation.

8.5 The estimate of the population mean from a stratified sample is: