~~Designation:E1402–99~~ Designation: E 1402 – 08

An American National Standard

~~Standard Terminology Relating to Sampling~~ Sampling Design[1]

This standard is issued under the fixed designation E 1402; 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

~~1.1This terminology covers those items related to statistical aspects of sampling.~~

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 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. This standard does not purport to address all of the safety concerns, if any, associated with its use.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

D 7430 Practice for Mechanical Sampling of Coal

E 105 Practice for Probability Sampling ~~of~~Of Materials

E 122 Practice for Calculating Sample Size to Estimate, With Specified Precision, the Average for a Characteristic of a Lot or Process

E 141 Practice for Acceptance of Evidence Based on the Results of Probability Sampling

E 456 Terminology Relating to Quality and Statistics

3. ~~Significance and Use~~

~~3.1This terminology standard is a subsidiary to Terminology E456.~~

~~3.2It provides definitions, descriptions, discussions, and comparison of terms.~~

~~4.~~Terminology ~~acceptance quality limit (AQL),~~

3.1 *Definitions:* Terminology E 456 contains a more extensive list of statistical terms.

3.1.1 *area sampling, n*— ~~quality level that is the worst tolerable process average when a continuing series of lots is submitted for acceptance sampling.~~

~~Discussion—This concept only applies when a sampling scheme with rules for switching and discontinuation such as in ISO 2859-1 or ISO 3951 is used. Although individual lots with quality as bad as the acceptance quality limit may be accepted with fairly high probability, the designation of an acceptance quality limit does not suggest that this is a desirable quality level. Sampling schemes found in international standards such as ISO 2859-1, with their rules for switching and discontinuation of sampling inspection, are designed to encourage suppliers to have process averages consistently better than AQL. Otherwise, there is a high risk that the inspection severity will be switched to tightened inspection, under which the criteria for lot acceptance become more demanding. Once on tightened inspection, unless action is taken to improve the process, it is very likely that the rule requiring discontinuation of sampling inspection pending such improvement will be invoked.~~

~~**cluster sampling,**~~ *—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.

~~[1] This terminology is under the jurisdiction of ASTM Committee E11 on Quality and Statistics and is the direct responsibility of Subcommittee E11.70 on Editorial/Terminology.~~

~~Current edition approved Nov. 16, 2001. Published February 2000. Originally published as E1402–91. Last previous edition E1402–96.~~

[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 Oct. 1, 2008. Published January 2009.

[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* ~~, Vol 14.02.~~volume information, refer to the standard's Document Summary page on the ASTM website.

3.1.2 *bulk sampling*, *n*— when the primary sampling unit comprises a bundle of elementary units or a group of subunits, the term cluster sampling may be applied.

DISCUSSION—Examples of cluster sampling are: selection of city blocks as primary sampling units; selection of a household as a cluster of people (of which only one may be interviewed); selection of a bundle of rods or pipe from a shipment; and selection, from a shipment of cartons that contain boxes or packages within them.

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

3.1.3 *cluster sampling*, *n*—a form of multi-phase sampling, in which there are only two phases. See **phase**.

draw, —*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 term used in sample selection. See **step**—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*. final sample,

3.1.7 *primary sampling unit, PSU*, *n*—sample obtained at the final stage of multi-stage sampling.

multi-stage sampling, nested sampling— sampling in which the sample is selected by stages, the sampling units at each stage being from the larger sampling units chosen at the previous stage.

NOTE1—Multi-stage sampling is different from multiple sampling. (see **acceptance sampling**).

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

DISCUSSION—This concept requires that the universe (or population) has been divided into a discrete set of sampling units or can be so divided in the process of selecting the sample. Examples are cartons of a lot or shipment, bales of wool or jute, and units created in moving a bulk material such as coal or sand. These units are designated as the primary sampling units, which may be subsampled at further stages of the sampling procedure.

probability sample, —*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 of which the sampling units have been selected by a chance process such that, at each step of selection, a specified probability of selection can be attached to each sampling unit available for selection.

DISCUSSION—These probabilities of selection need not be equal, Also, see Practice E105 in this volume.

proportional sampling, —*a sample in which the sampling units are selected by a chance process such that a 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 such that the proportion of the sampling units (usually, psu's) selected for the sample from each stratum is the same (except for possible rounding effects).

DISCUSSION—The procedure for proportional sampling is to select a sample from each stratum of a stratified universe (or population) such that (except for possible rounding effects):

n sub 1/*N* sub 1=*n* sub 2/*N* sub 2=*n* sub *g*/*N* sub *g*,

—a method of selection in stratified sampling such that the proportions of the sampling units (usually, PSU's) 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

where:

n sub = the sample size, and number of units in the population to be sampled.

iN

N sub = the stratum size for the *i*th stratum, number of units in the sample.

in

g Y_i = the number of strata

Size here refers to the number of sampling units (usually, psu's) in the sample and in the stratum. See simple random sample and probability sample for methods of selection within each stratum.

sampling—process of drawing or constituting a sample.

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

DISCUSSION—For the simple random sample case, $f = n/N$ where n is the sample size and N is the number of sampling units available. When $f > 0.10$, estimation of the precision of an estimator should take account of this magnitude of f.

sampling with replacement, *n*—a procedure used with some probability sampling plans 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.

sampling without replacement, *n*—a procedure in which a selected sampling unit is set aside for the sample, and a previously unselected unit is selected at each step (or draw) of the sample selection procedure.

DISCUSSION—Most samplings, including simple random sampling and stratified random sampling, are conducted by sampling without replacement. Computer methods have been developed for making the sample selections. See **step**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 $\bar{y}$ .stratified random sample, *n*—a sample that is selected independently within each stratum of a universe or population.

DISCUSSION—The sample selection within each stratum is usually a simple random sample, but probability sampling with unequal probabilities may be used, or systematic sampling may be used. Further, in order to optimize the sampling plan, the proportion of the sampling units selected for the sample in each stratum may or may not be the same from one stratum to another (optimization requires taking account of differing variances between the strata). Also, see **proportional sampling**.

subsample, *n*—sample taken from a sample of a population.

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 E 105 provides principles for designing probability sampling plans in relation to the objectives of study, costs, and practical constraints. Practice E 122 aids in specifying the required sample size. Practice E 141 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).

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 (e.g., (11)) that has been tested for equal digit frequencies and for serial independence, from a computer 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 ith item when the random number is less than $(n\text{-}a_i)/(N\text{-}i+1)$, 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

$$\bar{y} = \Sigma y_i/n \tag{1}$$

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

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

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

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

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

$$\bar{Y} = \sum_{i=1}^{N} Y_i/N \tag{4}$$

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

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

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 E 122 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:

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

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:

$$SE(\bar{X}\hat{R}) = \sqrt{\frac{1-f}{n}\Sigma\,(y_i - \hat{R}x_i)^2/(n-1)} \tag{7}$$

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 kth 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 E 141) for estimating sampling error if sample size n is large.

6.2 The sample average $\bar{y}$ is an unbiased estimate of the population mean. An estimate of the standard error of $\bar{y}$ is

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

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:

$$\bar{y}_{PPS} = \bar{z}\bar{X} \tag{8}$$

with standard error:

$$SE(\bar{y}_{PPS}) = \bar{X}\sqrt{1/n\sum_{i=1}^{n}(z_j - \bar{z})^2/(n-1)} \tag{9}$$

NOTE 2—It may be selected by the same method as was used in selecting the original sample, but need not be so.

NOTE 3—In sampling from bulk material, subsamples are often prepared by sample division. The subsample thus obtained is also called a "divided sample." See **sample division**.

systematic sampling, *n*—sample selection procedure in which every kth element is selected from the universe or population; for example, u, u+k, u+2k, u+3k, etc., where u is in the interval 1 to k.

DISCUSSION—If k=20 and u=7 is the initial unit selected, then sampling units, 7, 27, 47, 67, ..., would comprise the sample. When N/k is not an integer, there is a small bias due to the end effect. When u is selected by a chance process and N/k is an integer, the systematic sample will provide unbiased estimates of the population average or total. Situations for which N/k is not an integer usually ignore the small or negligible bias in estimating the mean or total. Schemes have been developed for non-integer N/k to overcome sampling bias.

Estimation of the precision of an average computed from a systematic sample is a difficult problem that has no generally satisfactory solution. Independent replicate systematic samples provide an approach to variance estimation, but have been rejected by some writers. In some ASTM situations where replicate samples may be obtained on a routine basis, the technique may be useful.