Designation: E1402 – 13

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.*

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 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

[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 Aug. 1, 2013. Published August 2013. Originally approved in 2008. Last previous edition approved in 2008 as E1402 – 08ε1. DOI: 10.1520/E1402-13.

[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.

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 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.

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

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,…,y_n. The estimate of the mean of the sampled population is

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

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).

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

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:

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

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 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:

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

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 $s^2 = np(1\text{-}p)/(n\text{-}1)$.

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:

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

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 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:

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

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 $C_i=\sum X_j$ 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:

$$\bar{y}_{PPS} = \bar{z}\bar{X} \quad (10)$$

with standard error:

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

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:

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

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.

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

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 $n_1, n_2,\ldots, n_L$ are called an allocation. Common allocations are:

(1) Proportional to N_h,

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

(3) Optimum, proportional to $N_hS_h/\sqrt{C_h}$ 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:

$$\bar{y}_{st} = \sum_{h=1}^{L}(N_h/N)\bar{y}_h \quad (14)$$

The estimated standard error of the mean is:

$$SE(\bar{y}_{st}) = \sqrt{\sum_{h=1}^{L}(N_h/N)^2[1-(n_h/N_h)]s_h{}^2/n_h} \quad (15)$$

8.6 Stratum divisions may be clear from the need to include various parts of the frame in the sample or from earlier surveys of the same type. If one has auxiliary information such as the

X_i of 5.6 then strata may usefully be constructed on the ordered frame. One recommended method uses the cumulative square root frequency (**16**). First form a grouped frequency distribution of the X_i having many small bins. Take the square root of each frequency and calculate their cumulative sums. Choose stratum boundaries so that they create approximately equal intervals on the cumulative square root of frequency scale. The number of strata, L, should be not more than six. Select the stratified sample and calculate following 8.2 – 8.5.

9. Quota Sampling

9.1 If strata can be defined with some rough knowledge of N_1, N_2, etc. then so-called quotas can be calculated proportional to the strata sizes. Enumerators can be instructed to use their own ways to reach addresses but required to determine stratum membership of prospective units to be selected and to ensure that the quotas are met but not exceeded. Enumerators usually prefer such schemes since human nature seems to prefer having control over selection. No general principles can be applied to estimate bias and precision for quota sampling. By conducting special experimental studies one can determine if the biases are tolerable (**17**).

10. Cluster Sampling

10.1 In cluster sampling, larger sized sampling units (clusters) are formed from smaller subunits. Clusters are sampled using a probability sampling method, and the subunits within each selected cluster are measured. There can be cost advantages to bringing all members of a cluster into a sample as a group. However, units belonging to the same cluster tend to be similar, which increases sampling variance over simple random sampling of the population of subunits for the same total number of subunits.

10.2 When clusters are of equal size, n clusters are selected by simple random sampling (SRS), and all items in selected clusters are measured, then the estimate of the overall average across units is

$$\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i \tag{16}$$

where $\bar{y}_i$ is the average over units in the i-th selected cluster. The estimated standard error of $\bar{\bar{y}}$ is based on the standard error of the average for simple random sampling.

$$SE\,(\bar{\bar{y}}) = \sqrt{\frac{1-f}{n}\sum_{i=1}^{n} (\bar{y}_i - \bar{\bar{y}})^2/(n-1)} \tag{17}$$

10.3 When clusters are of unequal size, n clusters are selected by simple random sampling (SRS), and all units in selected clusters are measured, then the estimate of the overall average across units is:

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{n} m_i \bar{y}_i}{\sum_{i=1}^{n} m_i} \tag{18}$$

where m_i is the number of units in the i-th sampled cluster and $\bar{y}_i$ is their average. The estimated standard error of $\bar{\bar{y}}$ is based on the estimated standard error of a ratio estimate (cluster size m_i takes the role of x_i , cluster total $m_i\bar{y}_i$ the role of y_i) for simple random sampling.

$$SE(\bar{\bar{y}}) = \sqrt{\frac{1-f}{n\bar{M}^2}\sum_{i=1}^{n} (m_i\bar{y}_i - m_i\bar{\bar{y}})^2} \tag{19}$$

where $\bar{M} = \sum_{i=1}^{N} M_i/N$ is the mean size of the population of clusters. If the number of units in clusters is only known for sampled clusters, then the mean size of sampled clusters $\bar{m} = \sum_{i=1}^{n} m_i/n$ may be substituted for $\bar{M}$ in Eq 19.

10.4 Choice of the correct size of cluster is a major design issue in cluster sampling. Just as for stratification and multistage sampling, costs are important, along with variances. Forming clusters is a clerically intensive task that oftentimes proceeds by trial and error and depends on auxiliary information on the frame. The U.S. Census Bureau clustered the nation's households into PSUs of similar size and cross-section and then clustered them into neighborhoods called area segments as SSUs prior to sampling them.

10.4.1 Consider variance first. If m elements were randomly selected to become a cluster then the variance of the cluster mean would be σ_1^2/m, where σ_1^2 is the variance of a single element. When elements adjacent to one another go into a cluster, the more realistic variance formula is $\sigma_1^2\, m^{-b}$. When elements of a cluster are similar to one another, b is less than 1. For example, when elements are foot of row in a wheat field and the variable is yield, b=0.7 has been found (**18**). When elements are households and the variable is income, b=0.5. When elements are tobacco plants and the variable is disease status, b can go as low as 0.1 for contagious diseases such as tobacco mosaic virus (**19**).

10.4.2 Next consider costs of getting to a cluster and the costs of enumerating elements within the cluster. The cost function commonly used is $C_T = C_0 + C_1 n + C_2 nm$ in which C_1 is the cost to sample a cluster and C_2 the cost for each element within a cluster.

10.4.3 The optimum cluster size is found by substituting for n from the cost function into the variance function $\sigma_1^2 m^{-b}/n$, and setting the derivative with respect to m equal to zero. This gives $m_{opt} = bC_1/[(b-1)C_2]$. The form of the result shows that the balance between C_1/C_2 and either b/(b-1) or the ratio of within variance to between variance governs the optimum cluster size. This result may look sharp but, in fact, there is a "flat" optimum and generally any value within 20 % of the optimum is an acceptable cluster size.

11. Multi-Stage Sampling

11.1 For two-stage sampling the frame is indexed by two integers as (i_1,i_2). The first tells what is the primary sampling unit (PSU) and the second tells the nested secondary sampling unit (SSU) number for each member of the population. PSUs are indexed from $i_1=1$ to $i_1=N_1$. If the frame is balanced then every PSU contains the same number, N_2, of SSUs. This notation can be extended to tertiary stage units (TSUs) indexed by $i_3=1, 2, \ldots, N_3$, and to i_4, i_5, etc.

11.1.1 When the nested frame has N_1 largest units (PSU) all having N_2 next-largest units (SSU) all having N_3 third stage

units (TSU) and so on, then the frame is said to be balanced. A (random) selection is made of n_1 from the N_1 and then from each selected PSU a (random) selection is made of n_2 of the N_2 and from the selected SSUs a (random) selection is made of n_3 TSUs, and so on. Data are collected on the final stage units.

11.2 Sampling fractions in balanced multistage sampling are $f_1=n_1/N_1$ for the first stage, $f_2=n_2/N_2$ for the second stage, and so on.

11.3 When the frame is balanced, the estimate of the population mean, denoted $\bar{\bar{y}}$, is simply an overall unweighted average. To determine the sampling variance of $\bar{\bar{y}}$, calculate variances for each stage, pooling over previous stages. Thus, for two stage sampling,

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (\bar{y}_i - \bar{\bar{y}})^2}{n_1 - 1} \tag{20}$$

$$s_2^2 = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} (y_{ij} - \bar{y}_i)^2}{n_1 (n_2 - 1)}$$

The estimated standard error of $\bar{\bar{y}}$ is:

$$SE(\bar{\bar{y}}) = \sqrt{\frac{1 - f_1}{n_1} s_1^2 + \frac{f_1 (1 - f_2)}{n_1 n_2} s_2^2} \tag{21}$$

11.4 Multistage sampling designs used in practice frequently incorporate stratification or unequal probability sampling (for example, PPS), especially at the first stage that often contributes most to sampling variance.

12. Summary of Discrete Population Sampling Designs

12.1 The following table classifying discrete population sampling designs illustrates the differences among stratified, cluster, and multistage sampling and clarifies the terminology. For all methods, units to be sampled are classified into groups. Groups, and the units that are members of each selected group, may be either enumerated or subject to probability sampling.

		Members of Groups	
		All Selected	Sampled
Groups	All Selected		Stratified Sampling Group = stratum Member = sampling unit
	Sampled	Cluster Sampling Group = cluster Member = subunit	Two Stage Sampling Group = PSU Member = SSU

12.2 Auxiliary information for the population, such as a measure of size of the sampling unit or a preliminary estimate of the quantity of interest, can be used in more than one way to improve the precision of sampling. The variable may be used in a ratio estimate, as a basis for sampling probabilities, or as a basis for stratification or clustering.

12.3 If sampling a variable of interest for the first time, especially when the sample is to be large, a pilot sample, chosen for convenience and for coverage of a variety of sampling units, should be run, and the choice of design then based on comparisons among the standard error formulas. Such a pilot operation should also be used to observe enumerators in order to plan future training, to improve measuring instruments, and to collect cost information in order to optimize stratification or cluster size.

13. Area Sampling

13.1 Area sampling, or area frame sampling, was originally devised for estimating crop production on farm fields and evaluating social characteristics of households. It now has wide application including sampling regions of the earth's surface.

13.2 The variable of interest will be some characteristic that can be measured conveniently on a very small element of area (a square meter say) or on a very small enumeration element (a household). The sampling unit, called an area segment, is a piece of land having a clearly defined boundary and containing a compact cluster of these elements. The variable's value is recorded as an amount, often a summed amount, found in the area segment.

13.3 Area sampling effectively segments the population region into N area segments which can be sampled by simple random sampling or by systematic sampling or, more commonly, by creating equal-sized strata, called zones, and selecting two area segments in each zone (**20**).

13.4 Since it would be prohibitively expensive to segment the entire population, topographic maps, aerial photographs, or satellite images are used to define natural and political boundaries that form a preliminary subdivision into fairly large count units – so called because there must be some "count" data for each unit that is easily available and roughly related to the amounts of the variable of interest to be found in each such count unit.

13.5 Sizes in surface area come directly from the maps and sizes data on auxiliary variables related to the amount of the variable of interest come from population and economic censuses and directories. The count units are assigned numbers of area segments where the numbers are determined by the sizes. The arithmetic steps are the following:

13.5.1 List count units, with their sizes, in serpentine order so that adjacency on the list implies adjacency on the ground.

13.5.2 Cumulate the sizes.

13.5.3 Decide on some target size of segment depending (*1*) on the cost of getting to a segment versus the cost of enumerating per element within the segment and (*2*) on the increase in variance caused by clustering the elements in an area segment versus taking a sample of the same number of elements randomly from the population of all elements.

13.5.4 Divide total size by target size and get a provisional total number of area segment sampling units (SU). Round this provisional number to a nearby integer having many divisors (such as a multiple of 200) to be N. Final target segment size is total size divided by this N. Divide the cumulative sizes by final segment size and round to get cumulative sampling unit (CSU) numbers. Difference these CSU numbers to assign definite numbers of sampling units to the count units.

13.5.5 When cumulative numbers of sampling units are repeated in the listing then the count units are merged into a single "finalized" count unit.

13.6 Treat the 1 to N listing of SUs as the sampling frame and apply a probability sampling design to the frame. For the two-per-zone design, suppose that n has been chosen, then create n/2 zones each of 2N/n sampling units (Recall that N was chosen to have divisors.) and randomly select two SUs in each zone.

13.7 Only those count units within which a sampling unit falls will be segmented into the assigned number of area segments. If only one segment is assigned then the existing (finalized) count unit boundary suffices to inform the enumerator where to work.

13.8 If more than one segment is assigned then a detailed map showing the geographic subdivision is required. Perhaps the segmenting can be done in the office from aerial photos but often an enumerator must go to the count unit and segment it from walking around it and noting internal features such as property lines, streams, power lines, etc. The aim of segmenting is to create subdivisions with approximately equal y-value amounts. Subjective and expert judgment needs to be used here.

13.9 After segmenting and numbering the created segments, the area segment to be enumerated must be chosen strictly by random number. Thus random numbers govern the entire selection procedure and subjective judgment is used only to reduce variance since it cannot create selection bias.

13.10 When all sampling units have been chosen with equal probabilities, as will normally be the case, the estimate of a population total will be the sample average over the n summed amounts multiplied by N and the standard error can be calculated from the formula appropriate to the design. Formulas (Eq 16) and (Eq 17) apply.

13.11 In the case of the two-per-zone design, (Eq 15) becomes:

$$SE(\bar{y}_{st}) = \sqrt{\left(1 - \frac{n}{N}\right) \sum_{h=1}^{n/2} d_h{}^2/n^2} \qquad (22)$$

where d_h is the difference between the summed y-values from the two sampling units in zone h. The standard error for the estimated total $N\bar{y}_{st}$ is N times Eq 22.

14. Bulk Sampling

14.1 Bulk sampling is a procedure for taking a portion of a lot of bulk material. The goal is for the sample portion to incorporate material from widely separated parts, and thus be representative of the whole, of the lot. Some useful references for bulk sampling are (**21-26**). Bulk sampling is done in stages of nested subsamples.

14.2 Bulk sampling is currently done by automatic and specialized machinery. A lot is better sampled off a conveyor belt than as a static mass. This often means extracting an increment at the point where the stream falls from one belt onto another or into a container. The device doing the sampling is called a cutter. It may be a slot in a moving plate or a slot under the belt or a collection container moving through the stream. The sampled material may itself then form a stream to be further sub-sampled by a secondary cutter. Mechanical sampling systems often include crushing between sub-samplings. Once sampled mass is lowered enough to be handled manually it is further crushed and poured through rifflers (slotted or rotary) that cause particles adjacent in the stream to be forced into different sub-sample portions or splits.

14.3 A bulk sampling design consists of specification of (*1*) lot size, (*2*) sub-sampling stages, (*3*) increments sizes, (*4*) numbers of increments, (*5*) particle top sizes and crushing used to attain them, (*6*) riffle splittings, (*7*) interleaved and sub-increment replications, (*8*) measurement methods to apply to test portion.

14.4 Nomenclature and specifications of a bulk sampling plan.

14.4.1 Although frames for listing increments and riffler splits can form part of the sampling protocol, they play a minor role. The unit that serves as sampling unit is the particle and it cannot ordinarily be listed.

14.4.2 Bulk sampling is done in stages whereby a first stage (i=1) sample is drawn and then sub-sampled to become a second stage sample, which is further sub-sampled. At the final stage (i=F) a test portion is sub-sampled to be analyzed.

14.4.3 At each stage there is an amount of material called the composite that leaves at that stage for the next. Its mass will be denoted M_i. In particular M_0 is lot mass and M_F is mass of the test portion. (Typically in coal sampling $M_0=10^{11}$ grams and $M_F=1$ gram.)

14.4.4 At each sampling stage the composite is made up of a number, n_i, of increments of average (targeted) mass m_i, so that $M_i=n_im_i$. After stages of this increment sampling come sample preparation stages in which riffle splitting takes place, until at some point the material is placed in a container and sent, as a laboratory sample, to be divided down, pulverized, subsampled usually by spatula scoop, and analyzed.

14.5 A routine bulk sampling plan may result in just one estimate and thus offers no basis for calculating sampling variance.

14.6 Many sampling systems can be run with replicates as follows (**27**). At the first stage of increment extraction the system takes $2n_1$ increments – double the usual number. These increments are numbered and the odd numbered ones go into an A gross sample while the even numbered ones are put into a B gross sample. Each gross sample is then processed to analysis in accord with the routine method. The difference between the A-B pair of analysis values estimates sampling error. Combining these estimates over a series of lots provides an estimate of sampling variance. The standard error for a sample is estimated by:

$$SE(y) = \sqrt{\sum_{k=1}^{K} (y_{kA} - y_{kB})^2/(2K)} \qquad (23)$$

where k indexes the K lots and the y-values are the analysis values from the two gross samples.

14.7 Many high-throughput mechanical sampling systems do not form the gross sample so the A-B replicates scheme will not serve. In these cases it may be possible to estimate sampling variance from the results of a bias test, which is an

important experiment in its own right to understand and to conduct before putting the mechanical system into operation. The sampling system is run for a short stretch of material, a batch, during which reference increments are extracted from the stopped belt. Practice D7430, Part D, has the details for coal sampling.

15. Keywords

15.1 area sample; bulk sample; cluster sample; finite population; probability proportional to size; quota sample; sample size; sampling frame; sampling without replacement; simple random sample; stratified sample

REFERENCES

(1) Barnett, V., *Sample Survey Principles and Methods*, Oxford U. Press, New York, NY, 1991.

(2) Cochran, W.G., *Sampling Techniques*, 3rd Ed., Wiley, New York, NY, 1977.

(3) Deming, W.E., *Some Theory of Sampling*, Dover, New York, NY, 1950.

(4) Deming, W.E., *Sample Design in Business Research*, Wiley, New York, NY, 1960.

(5) Hansen, M.H., Hurwitz, W.N., and Madow, W.G., *Sample Survey Methods and Theory*, 2 Vols, Wiley, New York, NY, 1953.

(6) Kish, L., *Survey Sampling*, Wiley, New York, NY, 1965.

(7) Lohr, S.L., *Sampling: Design and Analysis*, Duxbury Press, New York, NY, 1999.

(8) Sarndal, C.E., Swensson, B., and Wretman, J., *Model Assisted Survey Sampling*, Springer-Verlag, New York, NY, 1991.

(9) Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., and Asok, C., *Sampling Theory of Surveys with Applications*, 3rd Ed., Iowa State U. Press, 1984.

(10) Yates, F., *Sampling for Censuses and Surveys*, 4th Ed., Charles Griffin, London, 1981.

(11) Rand Corporation, *A Million Random Digits*, Free Press, 1955

(12) Brewer, K.R.W., and Hanif, M., *Sampling with Unequal Probabilities,* Springer-Verlag, New York, NY, 1983.

(13) Horvitz, D.G., and Thompson, D.J., "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, Vol 47, 1952, pp. 663– 685.

(14) Hartley, H.O., and Rao, J.N.K., "Sampling with Unequal Probabilities and Without Replacement," *Annals of Mathematical Statistics*, Vol 33 1968, pp. 350– 374.

(15) Neyman, J., "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Sampling," *Journal of the Royal Statistical Society*, Vol 97, 1934, pp. 558– 606.

(16) Dalenius, T., and Hodges, J.L., "Minimum variance stratification," *Journal of the American Statistical Association*, Vol 54, 1959, pp. 88–101.

(17) Moser, C.A., and Stuart, A., "An Experimental Study of Quota Sampling," *Journal of the Royal Statistical Society, Series A*, Vol 116, 1953, pp. 349 – 405.

(18) Smith, H.F., "An Empirical Law Describing Heterogeneity in the Yields of Agricultural Crops," *Journal of Agricultural Science*, Vol 28, 1937, pp. 1–23.

(19) Proctor, C.H., "Fitting H. F. Smith's Empirical Law to Cluster Variances for Use in Designing Multi-Stage Sample Surveys," *Journal of the American Statistical Association*, Vol 80, 1985, pp. 294–300; "Correction," Vol 86, 1991, p. 837.

(20) Monroe, J., and Finkner, A., *Handbook of Area Sampling*, Chilton Publishing Company, New York, NY, 1959.

(21) Bicking, C.A., "The Sampling of Bulk Materials," *Materials Research and Standards*, Vol 7, 1967, pp. 95– 116.

(22) Duncan, A.J., "Bulk Sampling: Problems and Lines of Attack," *Technometrics*, Vol 4, 1962, pp. 319 – 344.

(23) Ishikawa, K., "Some Experimental Methods for Bulk Material Sampling," *Report on Seminar on Sampling of Bulk Materials*, U.S. Japan Cooperative Science Program, National Science Foundation and Japan Society for Promotion of Science, Tokyo, 1965, pp. 187 –223.

(24) Pearson, E.S., "Sampling Problems in Industry," *Journal of the Royal Statistical Society, Series B*, Vol 1, 1934, pp. 107 – 153.

(25) *Symposium on Bulk Sampling, ASTM STP 114*, ASTM International, 1951.

(26) *Symposium on Coal Sampling, ASTM STP 162*, ASTM International, 1954.

(27) Gy, P.M., *Sampling of Particulate Materials: Theory and Practice*, Elsevier, New York, NY, 1982.