Designation: D 6582 – 00

Standard Guide for Ranked Set Sampling: Efficient Estimation of a Mean Concentration in Environmental Sampling[1]

This standard is issued under the fixed designation D 6582; 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 describes ranked set sampling, discusses its relative advantages over simple random sampling, and provides examples of potential applications in environmental sampling.

1.2 Ranked set sampling is useful and cost-effective when there is an auxiliary variable, which can be inexpensively measured relative to the primary variable, and when the auxiliary variable has correlation with the primary variable. The resultant estimation of the mean concentration is unbiased, more precise than simple random sampling, and more representative of the population under a wide variety of conditions.

1.3 *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 and health practices and determine the applicability of regulatory limitations prior to use.*

2. Referenced Documents

2.1 *ASTM Standards:*

D 5792 Practice for Generation of Environmental Data Related to Waste Management Activities: Development of Data Quality Objectives[2]

D 6044 Guide for Representative Sampling for Management of Waste and Contaminated Media[2]

3. Terminology

3.1 *Definitions:*

3.1.1 *auxiliary variable*, *n*—the secondary characteristic or measurement of interest.

3.1.1.1 *Discussion*—In ranked set sampling, information contained in an auxiliary variable is useful for ranking the samples. This ranking may mimic the rankings of the samples with respect to the values of the primary variable when there is correlation between the auxiliary variable and the primary variable. Auxiliary information may include visual inspection, inexpensive quick measurement, knowledge of operational history, previous site data, or any other similar information.

3.1.2 *data quality objectives (DQO) process*, *n*—a quality management tool based on the scientific method and developed by the U.S. Environmental Protection Agency (EPA) to facilitate the planning of environmental data collection activities. **(D 5792)**

3.1.3 *equal allocation*, *n*—this occurs when the number of sets in ranked set sampling is an integer multiple of the size of the set.

3.1.4 *primary variable*, *n*—the primary characteristic or measurement of interest.

3.1.5 *ranked set sampling*, *n*—a sampling method in which samples are ranked by the use of auxiliary information on the samples and only a subset of the samples are selected for the measurement of the primary variable.

3.1.6 *representative sample*, *n*—a sample collected in such a manner that it reflects one or more characteristics of interest (as defined by the project objectives) of a population from which it is collected. **(D 6044)**

3.1.6.1 *Discussion*—A representative sample can be a single sample, a collection of samples, or one or more composite samples. A single sample can be representative only when the population is highly homogeneous. **(D 6044)**

4. Significance and Use

4.1 Ranked set sampling is cost-effective, unbiased, more precise and more representative of the population than simple random sampling under a variety of conditions (**1**).[3]

4.2 Ranked set sampling (RSS) can be used when:

4.2.1 The population is likely to have stratification in concentrations of contaminant.

4.2.2 There is an auxiliary variable.

4.2.3 The auxiliary variable has strong correlation with the primary variable.

4.2.4 The auxiliary variable is either quick or inexpensive to measure, relative to the primary variable.

4.3 This guide provides a ranked set sampling method only under the rule of equal allocation. This guide is intended for

[1] This guide is under the jurisdiction of ASTM Committee D34 on Waste Management and is the direct responsibility of Subcommittee D34.01.01 on Planning for Sampling.

Current edition approved Aug. 10, 2000. Published October 2000.

[2] *Annual Book of ASTM Standards*, Vol 11.04.

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

those who manage, design, and implement sampling and analysis plans for management of wastes and contaminated media. This guide can be used in conjunction with the DQO process (see Practice D 5792).

5. Ranked Set Sampling (RSS)

5.1 Environmental sampling typically requires the identification of the locations where the samples are to collected. Subsequent analyses of these samples to quantify the characteristics of interest allow inference on the population mean concentration from the sample data.

5.2 A simple random sampling (SRS) approach is one sampling design that can be used. In this case, a set of random samples is identified and collected from a population and all of these samples are analyzed (for the primary variable).

5.3 Ranked set sampling (RSS) is similar to SRS in the identification and collection of the samples, but only a subset of the samples are selected for analysis. The selection is done by ranking the samples using auxiliary information on the samples and selecting a subset based on the rankings of the samples.

5.4 As can be seen from the steps described below, RSS is in fact a "stratified random sampling at the sample level," meaning that stratification of the population is induced after sampling and no construction of the strata is needed before sampling. Increased precision of stratified random sampling in the estimation of the population mean, relative to SRS, is well known, especially when the population is stratified by concentrations.

5.5 Increased precision of RSS relative to SRS means that the same precision can be achieved with fewer samples analyzed for the primary variable under RSS. RSS is therefore more cost-effective than SRS. When the objective is to minimize sampling and analytical costs, the number of samples can be determined so that RSS has precision equal to that of SRS at a lower cost.

5.6 The actual steps to conduct RSS are given below.

5.7 *Steps in Ranked Set Sampling (RSS)*:

5.7.1 Determine the total number of sample analyses (n) agreed to by the stakeholders. A planning process, such as a data quality objectives (DQO) process (Practice D 5792), may be used to determine this number.

5.7.2 Determine the primary variable and the auxiliary variable of interest.

5.7.3 Determine the size of the set, m. Study the auxiliary measurement and determine its capability in ranking the samples. For example, if the auxiliary measurement is visual inspection, its capability in ranking the samples may be somewhat limited. Namely, it may be capable of ranking 3–4 samples, but may have difficulty in ranking greater than 5 or 6 samples based on visual inspection; thus, the preferred size of the set (m) in ranked set sampling is about 3 or 4. On the other hand, an instrument-based quick-test may be capable of a larger m (see 5.14 for ranking criteria).

5.7.4 Calculate the needed number of replicates, r (the number of times the ranked sets are to be repeated). Divide n by m and round it up to whole number to obtain the needed r. Namely, $r = n/m$ and round up to whole number.

5.7.5 Randomly select a total of m^2r samples from the population, for example, by simple random sampling design, and randomly divide them into r replicates, with m^2 (m times m) samples in each replicate.

NOTE 1—In practice, the m^2r samples may not be taken all at once. More often, m random samples may be taken from a geographical sub-area of the population and are then ranked according to the auxiliary variable. This is repeated m times to obtain the first replicate of m^2 (m times m) samples. This entire process is repeated r times to obtain the needed r replicates.

5.7.6 Start with the first replicate of m^2 (m times m) samples. Arrange these samples into m sets of size m (an m by m matrix).

5.7.7 For each of the m sets in this replicate, rank the samples within each set by using the auxiliary measurement on the samples. When the observations on the auxiliary variable cannot be distinguished from each other, these observations are called "ties." Ties can be broken arbitrarily (namely, arbitrarily assigning one rank to one sample and a succeeding rank to the other).

5.7.8 Select samples for the measurement on the primary variable as follows. In set i, select and measure the sample with rank i, $i = 1, 2, ..., m$. Completion of this step leads to a total of m samples to be analyzed for the primary variable, out of a total of m^2 samples collected.

5.7.9 Repeat steps 5.7.6 through 5.7.8 for r times to obtain a total of $m \times r = n$ samples to be analyzed and measured for the primary variable.

5.8 Since the number of sets (m) in step 5.7.6 equals the size of the set (m), this is called equal allocation. RSS under unequal allocation tends to have additional gains in precision, relative to equal allocation; but, this gain is, in general, not large compared to the gain against SRS, and is not covered in this guide.

5.9 The value of n can be the total number of samples for which the budget can afford to analyze.

5.10 The rounding up in step 5.7.4 may cause the total number of analyses for the primary variable to exceed n. When this is the case, there are two options:

5.10.1 Obtain buy-in from the stakeholders to accept the slightly higher total number of sample analyses, or

5.10.2 Try different values of m and r to get the total number of analyses as close to n as possible.

5.11 *Estimation of Mean and Standard Error of the Mean*:

5.11.1 In 5.7, if $n = 12$, $m = 3$, and $r = 4$, the data on the primary variable obtained from the steps in that section may be summarized as in Table 1. The true mean concentration of the characteristic of interest is estimated by the arithmetic sample mean of the measured samples. For the hypothetical example in Table 1 (and assuming normal distribution of the data), the mean (M) is estimated as follows:

$$M = (X_{11} + X_{12} + X_{13} + X_{14} + X_{21} + X_{22} + ... + X_{34})/12 \quad (1)$$

The standard error of the mean (S_M) is estimated as follows:

$$S_M = \{[(X_{11} - X_{1.})^2 + (X_{12} - X_{1.})^2 + (X_{13} - X_{1.})^2 + (X_{14} - X_{1.})^2 + (X_{21} - X_{2.})^2 + (X_{22} - X_{2.})^2 + (X_{23} - X_{2.})^2 + ... + (X_{34} - X_{3.})^2]/(m^2r(r-1)\}^{1/2} \quad (2)$$