United States Environmental Protection Agency
Office of Environmental Information
Washington, DC 20460
EPA/240/R-02/005
December 2002

# Guidance on Choosing a Sampling Design for Environmental Data Collection

## for Use in Developing a Quality Assurance Project Plan

## EPA QA/G-5S

## FOREWORD

This document, *Guidance for Choosing a Sampling Design for Environmental Data Collection (EPA QA/G-5S)*, will provide assistance in developing an effective QA Project Plan as described in *Guidance for QA Project Plans (EPA QA/G-5)* (EPA 1998b). QA Project Plans are one component of EPA's Quality System. This guidance is different from most guidance in that it is not meant to be read in a linear or continuous fashion, but to be used as a resource or reference document. This guidance is a "tool-box" of statistical designs that can be examined for possible use as the QA Project Plan is being developed.

EPA works every day to produce quality information products. The information used in these products are based on Agency processes to produce quality data, such as the quality system described in this document. Therefore, implementation of the activities described in this document is consistent with EPA's Information Quality Guidelines and promotes the dissemination of quality technical, scientific, and policy information and decisions.

This document provides guidance to EPA program managers, analysts, and planning teams on statistically based sampling schemes. It does not impose legally binding requirements and the methods described may not apply to a particular situation based on the circumstances. The Agency retains the discretion to adopt approaches on a case-by-case basis that may differ from the techniques described in this guidance. EPA may periodically revise this guidance without public notice. It is the intent of the Quality Staff to revise the document to include: new techniques, corrections, and suggestions for alternative techniques. Future versions of this document will include examples in depth that illustrate the strengths of each statistical design.

This document is one of the *U.S. Environmental Protection Agency Quality System Series* documents. These documents describe the EPA policies and procedures for planning, implementing, and assessing the effectiveness of a Quality System. Questions regarding this document or other *Quality System Series* documents should be directed to the Quality Staff:

U.S. Environmental Protection Agency
Quality Staff (2811R)
1200 Pennsylvania Ave., NW
Washington, D.C. 20460
Phone: (202) 564-6830
Fax: (202) 565-2441
E-mail: quality@epa.gov

Copies of EPA *Quality System Series* documents may be obtained from the Quality Staff or by downloading them from epa.gov/quality/index.html.

## TABLE OF CONTENTS

## FIGURES

## TABLES

## BOXES

# CHAPTER 1

# INTRODUCTION

This document provides guidance on how to create sampling designs to collect environmental measurement data. This guidance describes several relevant basic and innovative sampling designs, and describes the process for deciding which design is right for a particular application.

## 1.1 WHY IS SELECTING AN APPROPRIATE SAMPLING DESIGN IMPORTANT?

The sampling design is a fundamental part of data collection for scientifically based decision making. A well-developed sampling design plays a critical role in ensuring that data are sufficient to draw the conclusions needed.[1] A sound, science-based decision is based on accurate information. To generating accurate information about the level of contamination in the environment, you should consider the following:

- the appropriateness and accuracy of the sample collection and handling method,
- the effect of measurement error,
- the quality and appropriateness of the laboratory analysis, and
- the representativeness of the data with respect to the objective of the study.

Of these issues, **representativeness** is addressed through the sampling design. Representativeness may be considered as the measure of the degree to which data accurately and precisely represent a characteristic of a population, parameter variations at a sampling point, a process condition, or an environmental condition [American National Standards Institute/American Society for Quality Control (ANSI/ASQC) 1994]. Developing a sampling design is a crucial step in collecting appropriate and defensible data that accurately represent the problem being investigated.

For illustration, consider Figure 1-1, a site map for a dry lagoon formerly fed by a pipe. Assuming that good field and laboratory practices are exercised and adequate quality control is implemented, the analytical results of soil samples drawn from randomly located sites A, B, and C may be representative if the objective is to address whether the pipe has released a particular contaminant. However, these data are not representative if the objective is to estimate the average concentration level of the entire old lagoon. For that estimation, random sampling locations should be generated from

[1]Note: Sampling design is not the only important component. The methods used in sample handling and extraction are equally important to the quality of the data. The United States Environmental Protection Agency produces extensive guidance on sampling methods and field sampling techniques for different regulations, regions, and programs that are not addressed in this document. In addition, measurement error affects the ability to draw conclusions from the data. *Guidance on Data Quality Indicators (QA/G-5i)* (EPA, 2001) contains information on this issue.

the entire site of the old lagoon (for example, perhaps including samples at D, E, and F). If a sampling design results in the collection of nonrepresentative data, even the highest quality laboratory analysis cannot compensate for the lack of representative data. The selection of the appropriate sampling design is necessary in order to have data that are representative of the problem being investigated.[2]

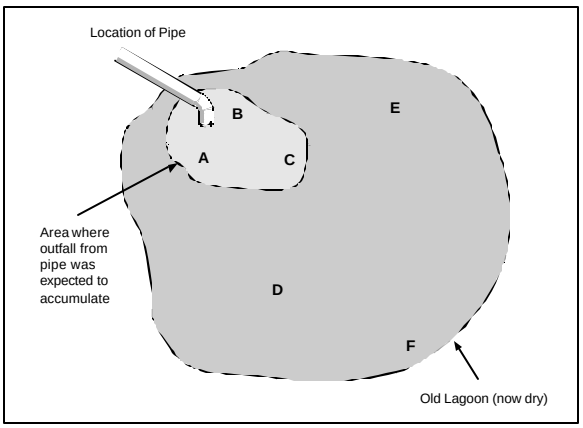


**Figure 1-1. Site Map for Old Lagoon**

This document provides technical guidance on specific sampling designs that can be used to improve the quality of environmental data collected. Based in statistical theory, each chapter explains the benefits and drawbacks of each design and describes relevant examples of environmental measurement applications. To choose a sampling design that adequately addresses the estimation or decision at hand, it is important to understand what relevant factors should be considered and how these factors affect the choice of an appropriate sampling design.

## 1.2 WHAT TYPES OF QUESTIONS WILL THIS GUIDANCE ADDRESS?

Often it is difficult in practice to know how to answer questions regarding how many samples to take and where they should be taken. The development of a sampling design will answer these questions after considering relevant issues, such as variability. Box 1-1 outlines the questions that are relevant to choosing a sampling design.

**Box 1-1. Questions that this Document Will Help to Address**

- What aspects of the problem should be considered for creating a sampling design?
- What are the types of designs that are commonly used in environmental sampling?
- What are some innovative designs that may improve the quality of the data?
- Which designs suit my problem?
- How should I design my sampling to provide the right information for my problem given a limited budget for sampling?
- How do I determine how much data are needed to make a good decision?

[2]Note: The problem of what constitutes "representativeness" is complex and further discussion may be found in *Guidance on Data Quality Indicators Peer Review Draft (QA/G-5i) (EPA, 2001)*.

## 1.3 WHO CAN BENEFIT FROM THIS DOCUMENT?

This document will be useful to anyone planning data collection from any type of environmental media including soil, sediment, dust, surface water, groundwater, air, vegetation, and sampling in indoor environments. The document contains information that will help those who are not extremely familiar with statistical concepts as well as those who are more comfortable with statistics. To this end, varying degrees of detail are provided on the various sampling designs, which should be used according to ability. The potential benefits for different types of users are shown in Table 1-1. This document is meant to apply to all environmental media; examples in this document provides information on innovative designs not discussed in earlier EPA documents.

The guidance document is designed for users who are not necessarily well versed in statistics. The document is written in plain language, and is designed to minimize technical jargon and provide useful explanations for those who might not already be familiar with the concepts described. In some chapters, more advanced material and more advanced references have been provided for statisticians - these have been marked as "more advanced."

**Table 1-1. Potential Benefits for Users**

| Potential User | Benefit to the User |
|---|---|
| **Environmental Scientist** or **Environmental Engineer** who is planning the sampling or **Project Manager** planning the investigation and reviewing the sampling plan | • An understanding of various sampling designs and the conditions under which these designs are appropriate<br>• An understanding of how sampling design affects the quality of the data and the ability to draw conclusions from the data<br>• An understanding of the appropriate uses of professional judgment<br>• The information needed to choose designs that may increase the quality of the data at the same cost as compared to typical sampling approaches (for example, Ranked Set Sampling) |
| **Risk Assessor** or **Data Analyst** who will be using the data | • An understanding of the advantages and limitations of data collected using various sampling designs<br>• The ability to draw scientifically based conclusions from data based on different types of designs<br>• The ability to match assessment tools to the sampling design used |
| **Statistician** assisting with the development and review of the sampling plan | • Tables, figures, and text that will help communicate important information about choosing a sampling design to colleagues working on the design who are not well versed in statistics<br>• Advanced references to support more complex design development |

## 1.4 HOW DOES THIS DOCUMENT FIT INTO THE EPA QUALITY SYSTEM?

Analysts should use systematic planning in order to collect data that will allow them to draw scientifically based conclusions. There are many cases in which data have been collected, but when the decision maker examines the data to draw conclusions, he or she finds that the data do not match the needs of the decision. Such problems can be avoided by using a systematic planning process to design the data collection. This process accounts for user's needs before the data are collected.

When data are being used in direct support of a decision. the Agency's recommended systematic plnning tool is the Data Quality Objectives (DQO) Process as described in EPA 2000b. For systematic planning of environmental data collection, EPA prefers the Data Quality Objectives (DQO) process described in the data quality objectives guidance (EPA, 2000b). A sampling design is chosen in Step 7 of the DQO Process based on the parameters specified in the other steps in the DQO Process. In this guidance, the activities of DQO Step 7 are explained in Chapter 3 (i.e., the process of choosing a sampling design), and a full discussion of the factors that should be considered in Step 7 of the DQO Process is given in Section 3.2.

Figure 1-2 illustrates the life-cycle of environmental data in the EPA Quality System. The process begins with systematic planning. Developing a sampling design is the last step in systematic planning, and is explained briefly in Step 7 of *Guidance for the Data Quality Objectives Process (QA/G-4)* (EPA, 2000b). This guidance document on sampling design is intended to expand greatly on the general details provided in that guidance. Information from the other steps in the systematic planning process are used as input to developing the sampling design. This process is described in detail in Chapter 3 of this guidance.

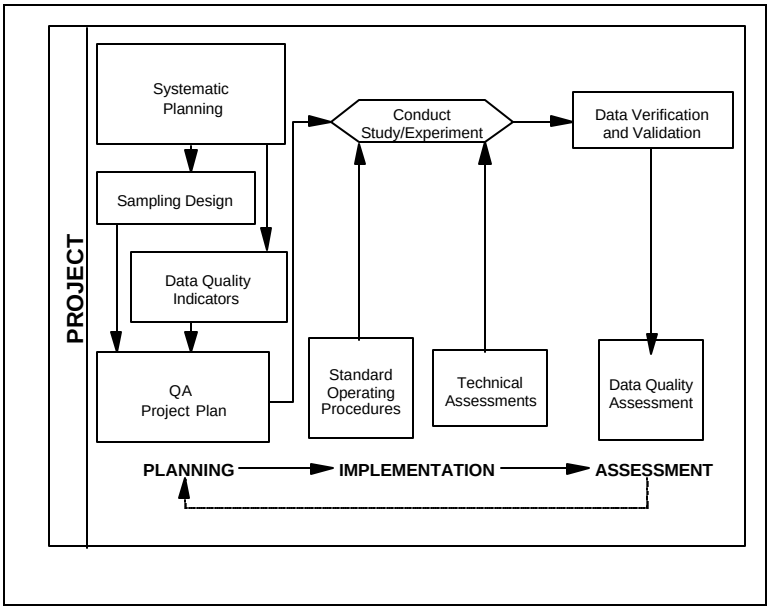


**Figure 1-2. Life-cycle of Data in the EPA Quality System**

**Data Quality Indicators** (DQIs) are specific calculations that measure performance as reflected in the DQOs and performance and acceptance criteria. DQIs include precision, accuracy, representativeness, completeness, consistency, and sensitivity, and are discussed at length in *Guidance on Data Quality Indicators (QA/G-5i)* (EPA, 2001). The choice of sampling design will have an impact on the DQIs. These indicators are addressed specifically for each project in the details of the Quality Assurance (QA) Project Plan.

The development of a sampling design is followed by the development of a QA Project Plan. A process for developing a QA Project Plan is described in *Guidance for Quality Assurance Project Plans (QA/G-5)* (EPA, 1998b).

After the QA Project Plan is developed and approved, data are collected during the study/experimental phase according to the plan. Quality is further assured by the use of standard operating procedures and audits (technical assessment). Finally, verification, validation, and quality assessment of the data complete the quality system data collection process.

## 1.5 WHAT SOFTWARE SUPPLEMENTS THIS GUIDANCE?

Visual Sampling Plan (VSP) is a software tool that contains some of the sampling plans discussed in this guidance. VSP supports the implementation of the DQO Process by visually displaying different sampling plans, linking them to the DQO Process, and determining the optimal sampling specifications to protect against potential decision errors. This easy-to-use program is highly visual, very graphic, and intended for use by non-statisticians. VSP may be obtained from http://dqo.pnl.gov.vsp.

## 1.6 WHAT ARE THE LIMITATIONS OR CAVEATS TO THIS DOCUMENT?

The scope of this document is limited to environmental measurement data. It does not explicitly address count data, survey (questionnaire) data, human exposure data, or experimental data collection, although some of the concepts described here are applicable to these types of studies. This guidance does not provide a complete catalogue of potential sampling designs used by EPA. These guidelines do not supercede regulatory requirements for specific types of sampling design, nor regional, state, or program guidance; rather, they are intended to supplement other guidance.

In addition, there are sampling designs that might be used in environmental data collection that are not discussed in this document. For example, double sampling, sequential sampling, quota sampling, and multi-stage sampling are all designs that are used for environmental data collection. Information on these designs can be found in other resources on sampling designs.

## 1.7 HOW IS THIS DOCUMENT ORGANIZED?

This document is designed to be used as a reference rather than be read from beginning to end. First-time users will probably want to skim Chapter 2 and read Chapter 3 before continuing to other chapters. Chapter 2 defines important concepts and terms, and introduces the types of sampling designs covered in this document, along with information on what specific types of situations call for which designs. Chapter 3 describes the process of developing a sampling design and discusses how input from a systematic planning process affects the choice of a sampling design.

The remaining chapters contain specific information about different sampling designs or protocols. Each chapter is formatted in a similar style to allow the reader to easily find information. A synopsis of the benefits and limitations of the design can be found in each chapter, so that readers can evaluate each design in light of their specific situation. Each chapter also contains at least one example and descriptions of applications of this design, where possible. Finally, each chapter has an appendix containing formulae and additional technical information.

Some designs are often used in conjunction with other designs; descriptions and examples of these types of studies are included. At the end of the document, a glossary defines key terms and a list of references contains citations for all referenced material and other materials used in developing this document.

The level of detail provided in the chapters varies based on the complexity of the design. For simpler designs, the chapter provides relatively complete information regarding how and when to implement this approach. For more complex designs, a general discussion is provided, along with references that can provide more information for the interested reader. It is assumed that a statistician would need to be involved in the development process for the more complex designs.

# CHAPTER 2

# OVERVIEW OF SAMPLING DESIGNS

## 2.1 OVERVIEW

### What does a sampling design consist of?

A complete sampling design indicates the number of samples and identifies the particular samples (for example, the geographic positions where these samples will be collected or the time points when samples will be collected). Along with this information, a complete sampling design will also include an explanation and justification for the number and the positions/timings of the samples. For a soil sample, the samples may be designated by longitude and latitude, or by measurements relative to an existing structure. For air or water measurements, the samples would be designated by longitude and latitude as well as by time. For example, for the measurement of particulates in air, a specified length of time would be set, such as 24 hours, in addition to the geographical location. The sampling design would note what time the air sample collection would begin (for example, 12:00 midnight on February 10, 2001), and when it would end (for example, 12:00 midnight on February 11, 2001). The measurement protocol would then specify when the sampler would be retrieved and how the sample would be analyzed.

### What is the purpose of a sampling design?

The goals of a sampling design can vary widely. Typical objectives of a sampling design for environmental data collection are:

- To support a decision about whether contamination levels exceed a threshold of unacceptable risk,
- To determine whether certain characteristics of two populations differ by some amount,
- To estimate the mean characteristics of a population or the proportion of a population that has certain characteristics of interest,
- To identify the location of "hot spots" (areas having high levels of contamination) or plume delineation,
- To characterize the nature and extent of contamination at a site, or
- To monitor trends in environmental conditions or indicators of health.

A well-planned sampling design is intended to ensure that resulting data are adequately representative of the target population and defensible for their intended use. Throughout the sampling design process, the efficient use of time, money, and human resources are critical considerations. A good design should meet the needs of the study with a minimum expenditure of resources. If resources

are limited or these are multiple objectives, tradeoffs may need to be made in the design. More information on how to go about doing this is contained in Chapter 3 on the sampling design process.

## 2.2 SAMPLING DESIGN CONCEPTS AND TERMS

Defining the population is an important step in developing a sampling plan. The **target population** is the set of all units that comprise the items of interest in a scientific study, that is, the population about which the decision maker wants to be able to draw conclusions. The **sampled population** is that part of the target population that is accessible and available for sampling. For example, the target population may be defined as surface soil in a residential yard, and the sampled population may be areas of soil in that yard not covered by structures or vegetation. Ideally, the sampled population and the target population are the same. If they are not, then professional judgment is used to verify that data drawn from the sampled population is appropriate for drawing conclusions about the target population.

A **sampling unit** is a member of the population that may be selected for sampling, such as individual trees, or a specific volume of air or water. It is important for study planners to be very specific when defining a sampling unit's characteristics with respect to space and time. A sampling unit should detail the specific components of a particular environmental media, for example, 10 cubic meters ($m^3$) of air passing through a filter located in downtown Houston on July 15, 2000. Some environmental studies have distinct sampling units such as trees, fish, or drums of waste material. However, such distinct sampling units may not be available in environmental studies requiring samples of soil, water, or other solid or liquid media. In this case, the sampling units are defined by the investigator and need to be appropriate for selecting a representative sample of material from the medium of interest. The physical definition of a sampling unit in terms of its "size, shape, and orientation" is referred to as the sample support (Starks, 1986). The **sampling frame** is a list of all the possible sampling units from which the sample can be selected. The **sample** is a collection of some of these sampling units.

**Sample support** represents that portion of the sampling unit, such as an area, volume, mass, or other quantity, that is extracted in the field and subjected to the measurement protocol (see definition below). It is a characteristic of a sample describing its relationship to the entity from which it was taken. It represents an area, mass, volume within the sampling unit. For example, if a sampling unit is a single tree, the sample support could be a core from the base of the tree. Or, if a sample unit is 10 grams of soil from a particular x-y coordinate, the sample support might be 1 gram of this soil after homogenization. Smaller sample support usually results in greater sampling variation (i.e., greater variability between sampling units) [see Section 21.5.3 of Pitard (1993)]. For example, soil cores with a 2-inch diameter and 6-inch depth usually have greater variability in contaminant concentrations than cores with a 2-inch diameter and 5-foot depth, much like composite samples have less variability than

individual specimens (see Chapter 9). Hence, the study objectives need to clearly define the sample support in order for the results (for example, sample mean and variance) to be clearly interpretable.

Once a sampling unit is selected, a **measurement protocol** is applied; a measurement protocol is a specific procedure for making observations or performing analyses to determine the characteristics of interest for each sampling unit. The measurement protocol would include the procedures for collecting a physical sample, handling and preparing the physical sample, applying an analytical method (including the sample preparation steps) to obtain a result (that is, to obtain the data for the sample), and protocol for resampling if necessary. If compositing of the samples is employed (so that measurements are made on the composites), then the measurement protocol would also include a **composite sampling protocol**, which indicates how many composites are to be formed, how many samples comprise each composite, and which samples are used to form each composite; the compositing protocol would also prescribe the compositing procedures (for example, for homogenization, for taking aliquots). The **sampling design** specifies the number, type, and location (spatial and/or temporal) of sampling units to be selected for measurement.

A water sampling example illustrates how these terms relate to one another. Consider a study designed to measure *E. coli* and *entercocci* levels in a specific swimming area of a lake. The target population is the water flowing through this area (delineated by buoys) from May 1 until September 15. The sampled population will be the water in the swimming area at 7 a.m. and 2 p.m at approximately 6 inches below the surface. The sampling units chosen for the study consist of 1-liter volumes of water at particular locations in the swimming area. In this case, the sample support is equal to the sampling unit, 1 liter of water. The measurement protocol calls for the use of a 2-liter beaker, held by a 6-inch handle. The sampler needs a nonmotorized boat (for example, a rowboat) to collect the sample so as to minimize the disturbance to the water. The sample is collected in the specified manner and poured into a 2-liter sample jar, up to the 1-liter line. The rest of the water in the beaker is discarded back into the lake. Each 1-liter container of water is taken to the lab for analysis within 6 hours and is analyzed according to current state standards. The sampling design calls for obtaining a minimum of two samples on each sampling day at 7 a.m. and 2 p.m or up to three times a day when there are indications of increased potential for contamination (for example, heavy rainfall). Sampling days are defined in the study and may be every day, every other day, or whatever frequency is appropriate for the particular problem at hand. The sampling design also specifies the exact locations where the samples should be drawn, which in this case were chosen at random.

Another important concept for sampling design is the **conceptual model**. At the outset of data collection activities, it is critical to develop an accurate conceptual model of the potential hazard. A conceptual model describes the expected source of the contaminant and the size and breadth of the area of concern, identifies the relevant environmental media and the relevant fate and transport pathways, and defines the potential exposure pathways. The model should also identify potential

sources of variability in the data (for example, inherent variability among sampling units in the population and variability associated with selecting and analyzing samples).

## 2.3 PROBABILISTIC AND JUDGMENTAL SAMPLING DESIGNS

There are two main categories of sampling designs: probability-based designs and judgmental designs. **Probability-based sampling designs** apply sampling theory and involve random selection of sampling units. An essential feature of a probability-based sample is that each member of the population from which the sample was selected has a known probability of selection. When a probability-based design is used, statistical inferences may be made about the sampled population from the data obtained from the sampling units. That is, when using a probabilistic design, inferences can be drawn about the sampled population, such as the concentration of fine particulate matter ($PM_{2.5}$) in ambient air in downtown Houston on a summer day, even though not every single "piece" of the downtown air is sampled. **Judgmental sampling designs** involve the selection of sampling units on the basis of expert knowledge or professional judgment.

Table 2-1 summarizes the main features of each main type of sampling design. Section 2.3.1 introduces judgmental sampling, and Chapter 4 contains more information on the benefits and limitations of this design. Sections 2.3.2 through 2.3.7 introduce the six probabilistic sampling designs, and Chapters 5 through 10 describe these in more detail. Reviewing these chapters will provide more details about the appropriate use of these designs.

**Table 2-1. Probability-based versus Judgmental Sampling Designs**

| | Probability-based | Judgmental |
|---|---|---|
| **Advantages** | • Provides ability to calculate uncertainty associated with estimates<br>• Provides reproducible results within uncertainty limits<br>• Provides ability to make statistical inferences<br>• Can handle decision error criteria | • Can be less expensive than probabilistic designs. Can be very efficient with knowledge of the site<br>• Easy to implement |
| **Disadvantages** | • Random locations may be difficult to locate<br>• An optimal design depends on an accurate conceptual model | • Depends upon expert knowledge<br>• Cannot reliably evaluate precision of estimates<br>• Depends on personal judgment to interpret data relative to study objectives |

Figure 2-1 illustrates the data collection process for both judgmental sampling and probabilistic sampling. Both processes start with defining the target population and the sampled population, and

each ends with data collection and analysis. The difference is seen when moving up the diagram, which shows how conclusions can be drawn about the sampled and target populations.

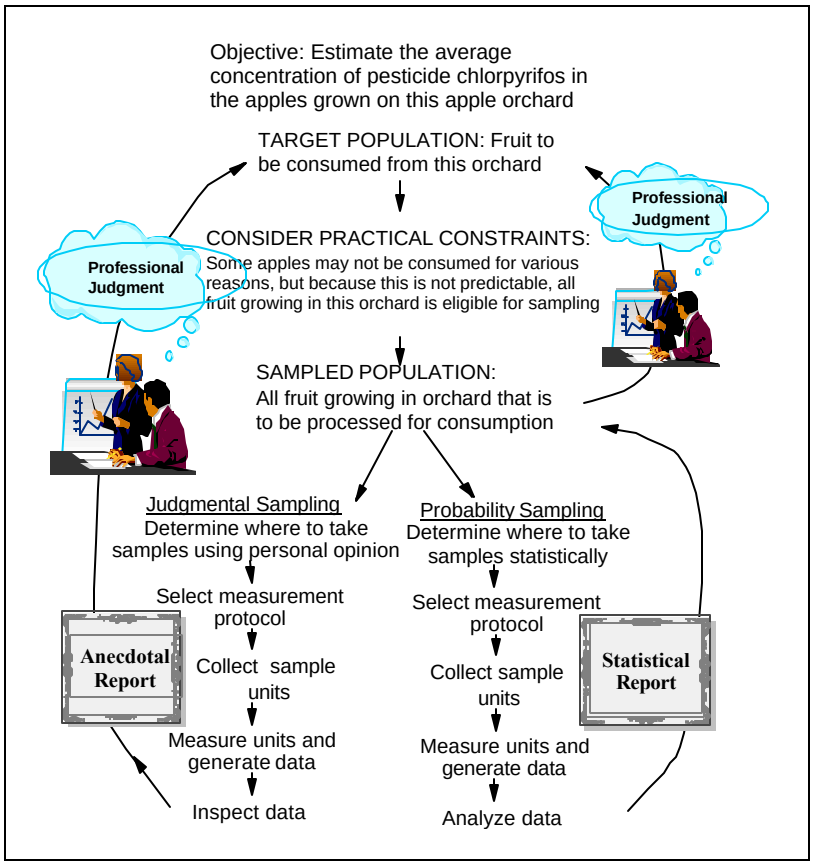


**Figure 2-1. Inferences Drawn from Judgmental versus Probabilistic Sampling Designs**

When using **probabilistic sampling**, the data analyst can draw quantitative conclusions about the sampled population. That is, in estimating a parameter (for example., the mean), the analyst can calculate a 95% confidence interval for the parameter of interest. If comparing this to a threshold, the analyst can state whether the data indicate that the concentration exceeds or is below the threshold with a certain level of confidence. Expert judgment is then used to draw conclusions about the target population based on the statistical findings about the sampled population. Expert judgment can also be used in other aspects of probabilistic sampling designs, such as defining strata in a stratified design. Such uses of expert judgment will be discussed in more detail in relevant sampling design chapters.

When using **judgmental sampling**, statistical analysis cannot be used to draw conclusions about the target population. Conclusions can only be drawn on the basis of professional judgment. The usefulness of judgmental sampling will depend on the study objectives, the study size and scope, and the degree of professional judgment available. When judgmental sampling is used, quantitative statements about the level of confidence in an estimate (such as confidence intervals) cannot be made.

## 2.4 TYPES OF SAMPLING DESIGNS

This guidance describes six sampling designs and one sampling protocol (i.e., composite sampling). Most of these designs are commonly used in environmental data collection. Some are designs that are not as commonly used but have great potential for improving the quality of

environmental data. Table 2-2 identifies the sampling designs discussed in this document, and indicates which chapter contains detailed information on each design. This section briefly describes each design, providing some information about the type of applications for which each design is especially appropriate and useful.

**Table 2-2. Sampling Designs Presented in this Guidance**

| Sampling Design/Protocol | Chapter | Use |
|---|---|---|
| Judgmental | 4 | Common |
| Simple Random | 5 | Common |
| Stratified | 6 | Common |
| Systematic and Grid | 7 | Common |
| Ranked Set | 8 | Innovative |
| Adaptive Cluster | 9 | Innovative |
| Composite | 10,11 | Common |

### 2.4.1 Judgmental Sampling

In judgmental sampling, the selection of sampling units (i.e., the number and location and/or timing of collecting samples) is based on knowledge of the feature or condition under investigation and on professional judgment. Judgmental sampling is distinguished from probability-based sampling in that inferences are based on professional judgment, not statistical scientific theory. Therefore, conclusions about the target population are limited and depend entirely on the validity and accuracy of professional judgment; probabilistic statements about parameters are not possible. As described in subsequent chapters, expert judgment may also be used in conjunction with other sampling designs to produce effective sampling for defensible decisions.

### 2.4.2 Simple Random Sampling

In simple random sampling, particular sampling units (for example, locations and/or times) are selected using random numbers, and all possible selections of a given number of units are equally likely. For example, a simple random sample of a set of drums can be taken by numbering all the drums and randomly selecting numbers from that list or by sampling an area by using pairs of random coordinates. This method is easy to understand, and the equations for determining sample size are relatively straightforward. An example is shown in Figure 2-2. This figure illustrates a possible simple random sample for a square area of soil. Simple random sampling is most useful when the population of interest is relatively homogeneous; i.e., no major patterns of contamination or "hot spots" are expected. The main advantages of this design are:

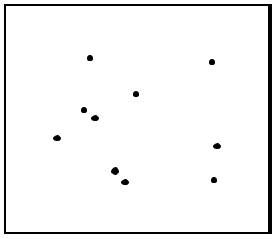

**Figure 2-2. Simple Random Sampling**

(1) It provides statistically unbiased estimates of the mean, proportions, and variability.
(2) It is easy to understand and easy to implement.
(3) Sample size calculations and data analysis are very straightforward.

In some cases, implementation of a simple random sample can be more difficult than some other types of designs (for example, grid samples) because of the difficulty of precisely identifying random geographic locations. Additionally, simple random sampling can be more costly than other plans if difficulties in obtaining samples due to location causes an expenditure of extra effort.

### 2.4.3 Stratified Sampling

In stratified sampling, the target population is separated into nonoverlapping strata, or subpopulations that are known or thought to be more homogeneous (relative to the environmental medium or the contaminant), so that there tends to be less variation among sampling units in the same stratum than among sampling units in different strata. Strata may be chosen on the basis of spatial or temporal proximity of the units, or on the basis of preexisting information or professional judgment about the site or process. Figure 2-3 depicts a site that was stratified on the basis of information about how the contaminant is present based on wind patterns and soil type and on the basis of surface soil texture. This design is useful for estimating a parameter when the target population is heterogeneous and the area can be subdivided based on expected contamination levels. Advantages of this sampling design are that it has potential for achieving greater precision in estimates of the mean and variance, and that it allows computation of reliable estimates for population subgroups of special interest. Greater precision can be obtained if the measurement of interest is strongly correlated with the variable used to make the strata.

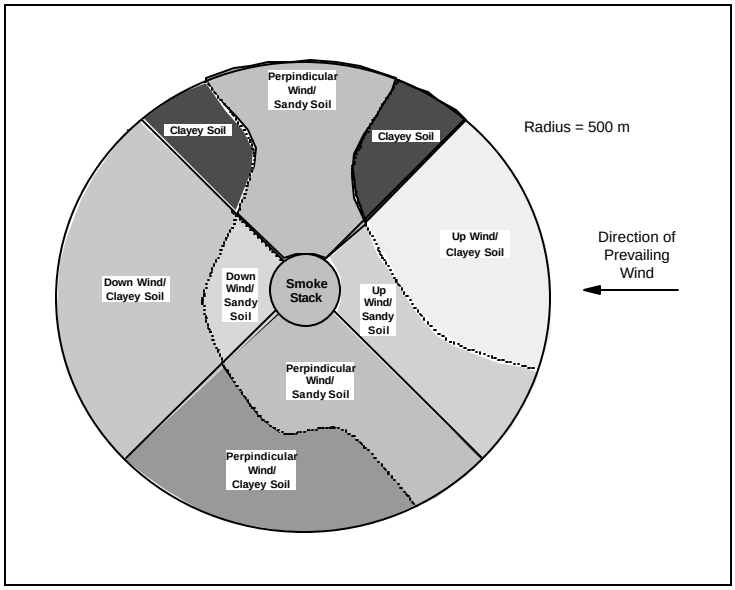


**Figure 2-3. Stratified Sampling**

### 2.4.4 Systematic and Grid Sampling

In systematic and grid sampling, samples are taken at regularly spaced intervals over space or time. An initial location or time is chosen at random, and then the remaining sampling locations are defined so that all locations are at regular intervals over an area (grid) or time (systematic). Examples

of systematic grids include square, rectangular, triangular, or radial grids [Section 16.6.2 of Myers (1997)].

In random systematic sampling, an initial sampling location (or time) is chosen at random and the remaining sampling sites are specified so that they are located according to a regular pattern (Cressie, 1993) for example, at the points identified by the intersection of each line in one of the grids shown in Figure 2-4. Systematic and grid sampling is used to search for hot spots and to infer means, percentiles, or other parameters and is also useful for estimating spatial patterns or trends over time. This design provides a practical and easy method for designating sample locations and ensures uniform coverage of a site, unit, or process.

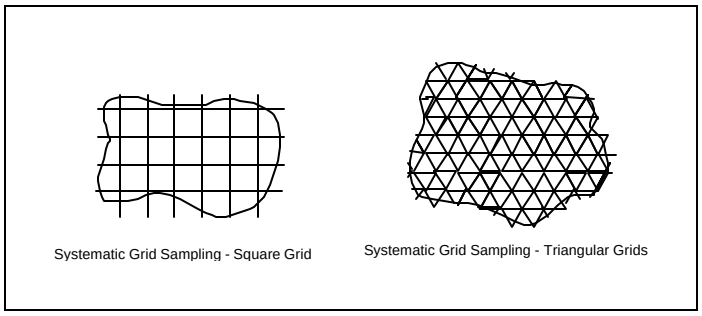


**Figure 2-4. Systematic/Grid Sampling**

### 2.4.5 Ranked Set Sampling

Ranked set sampling is an innovative design that can be highly useful and cost efficient in obtaining better estimates of mean concentration levels in soil and other environmental media by explicitly incorporating the professional judgment of a field investigator or a field screening measurement method to pick specific sampling locations in the field. Ranked set sampling uses a two-phase sampling design that identifies sets of field locations, utilizes inexpensive measurements to rank locations within each set, and then selects one location from each set for sampling.

In ranked set sampling, $m$ sets (each of size $r$) of field locations are identified using simple random sampling. The locations are ranked independently within each set using professional judgment or inexpensive, fast, or surrogate measurements. One sampling unit from each set is then selected (based on the observed ranks) for subsequent measurement using a more accurate and reliable (hence, more expensive) method for the contaminant of interest. Relative to simple random sampling, this design results in more representative samples and so leads to more precise estimates of the population parameters.

Ranked set sampling is useful when the cost of locating and ranking locations in the field is low compared to laboratory measurements. It is also appropriate when an inexpensive auxiliary variable (based on expert knowledge or measurement) is available to rank population units with respect to the variable of interest. To use this design effectively, it is important that the ranking method and analytical method are strongly correlated.

### 2.4.6 Adaptive Cluster Sampling

In adaptive cluster sampling, *n* samples are taken using simple random sampling, and additional samples are taken at locations where measurements exceed some threshold value. Several additional rounds of sampling and analysis may be needed. Adaptive cluster sampling tracks the selection probabilities for later phases of sampling so that an unbiased estimate of the population mean can be calculated despite oversampling of certain areas. An example application of adaptive cluster sampling is delineating the borders of a plume of contamination.

Initial and final adaptive sampling designs are shown in Figure 2-5. Initial measurements are made of randomly selected primary sampling units using simple random sampling (designated by squares in Figure 2-5). Whenever a sampling unit is found to show a characteristic of interest (for example, contaminant concentration of concern, ecological effect as indicated by the shaded areas in the figure), additional sampling units adjacent to the original unit are selected, and measurements are made.

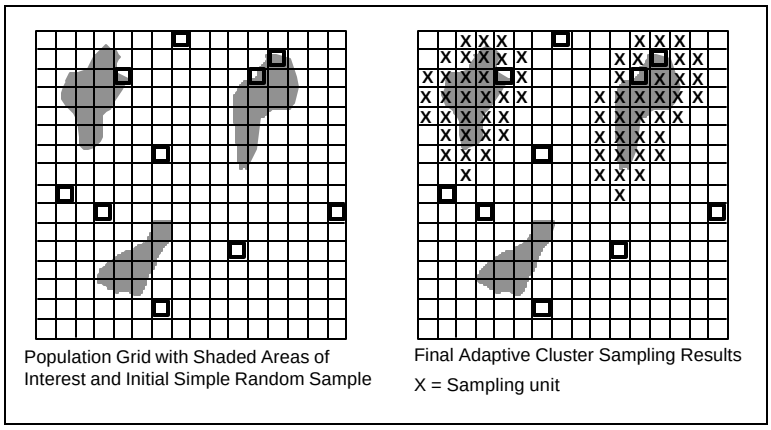


**Figure 2-5. Adaptive Cluster Sampling**

Adaptive sampling is useful for estimating or searching for rare characteristics in a population and is appropriate for inexpensive, rapid measurements. It enables delineating the boundaries of hot spots, while also using all data collected with appropriate weighting to give unbiased estimates of the population mean.

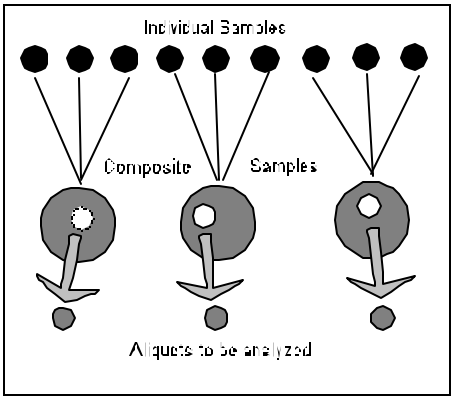


**Figure 2-6. Composite Sampling**

### 2.4.7 Composite Sampling

In composite sampling (illustrated in Figure 2-6), volumes of material from several of the selected sampling units are physically combined and mixed in an effort to form a single homogeneous sample, which is then analyzed. Compositing can be very cost effective because it reduces the number of chemical analyses needed. It is most cost effective when analysis costs are large relative to sampling costs; it demands, however, that there are no safety hazards or potential biases (for example, loss of volatile organic components) associated with the compositing process.

Compositing is often used in conjunction with other sampling designs when the goal is to estimate the population mean and when information on spatial or temporal variability is not needed. It can also be used to estimate the prevalence of a rare trait. If individual aliquots from samples comprising a composite can be retested on a new portion, retesting schemes can be combined with composite sampling protocols to identify individual units that have a certain trait or to determine those particular units with the highest contaminant levels.

# CHAPTER 3

# THE SAMPLING DESIGN PROCESS

## 3.1 OVERVIEW

### What are the objectives of the sampling design process?

The sampling design process should match the needs of the project with the resources available. The needs generally consist of the study objectives and the tolerable limits on uncertainty. The resources may include personnel, time, and availability of financial resources. The goal of the process is to use all of the information available so that the data collected meets the needs of the decision maker.

### Who is typically involved in the sampling design process?

The sampling design process typically includes a multi-disciplinary group (such as a DQO development team) that is involved in systematic planning at the beginning and at key review points. This team should include the decision maker or end user of the data. More rigorous technical activities will likely be performed by statisticians or by environmental scientists or engineers who have training and experience in environmental statistics.

## 3.2. INPUTS TO THE SAMPLING DESIGN PROCESS

### What outputs from the systematic planning process are incorporated into the sampling design process?

It is EPA policy (EPA, 2000c) that all EPA organizations use a systematic planning process to develop acceptance or performance criteria for the collection, evaluation, or use of environmental data. Systematic planning identifies the expected outcome of the project, the technical goals, the cost and schedule, and the acceptance criteria for the final result. The Data Quality Objectives (DQO) Process is the Agency's recommended planning process when data are being used to select between two opposing conditions, such as decision-making or determining compliance with a standard. The outputs of this planning process (the data quality objectives themselves) define the performance criteria. The DQO Process is a seven-step planning approach based on the scientific method that is used to prepare for data collection activities such as environmental monitoring efforts and research. It provides the criteria that a sampling design should satisfy, where to collect samples; tolerable decision error rates; and the number of samples to collect.

DQOs are qualitative and quantitative statements, developed in the first six steps of the DQO Process (Figure 3-1), that define the purpose for the data collection effort, clarify the kind of data needed, and specify the limits on decision errors needed for the study. These outputs are used in the final DQO step to develop a sampling design that meets the performance criteria and other design constraints. The DQO Process helps investigators ensure that the data collected are of the right type, quantity, and quality needed to answer research questions or support environmental decisions, and ensures that valuable resources are spent on collecting only those data necessary to support defensible decisions.

The DQO Process is a systematic planning approach for data collection that is based on the scientific method and uses a seven-step process. Although the DQO Process is typically described in linear terms, it is really a flexible process that relies on iteration and modification as the planning team works through each step, thus allowing early steps to be revised in light of information developed from subsequent steps.

**The Steps of the DQO Process**

**Step 1: State the Problem.** This step defines the problem clearly, identifies the primary decision maker and planning team members, and determines the available budget, personnel, and schedule deadlines.

**Step 2: Identify the Decision.** The key activities are to develop an appropriate decision statement: identify the principal study question, define alternative actions that could result from resolving the principal study question, link the principal study question to possible actions, and organize multiple decisions.

**Step 3: Identify the Inputs to the Decision.** These activities include identifying the type and sources of information needed to resolve the decision statement, identifying information needed to establish the action level, and confirming that suitable methods exist.

**Step 4: Define the Boundaries of the Study.** This step specifies the characteristics that

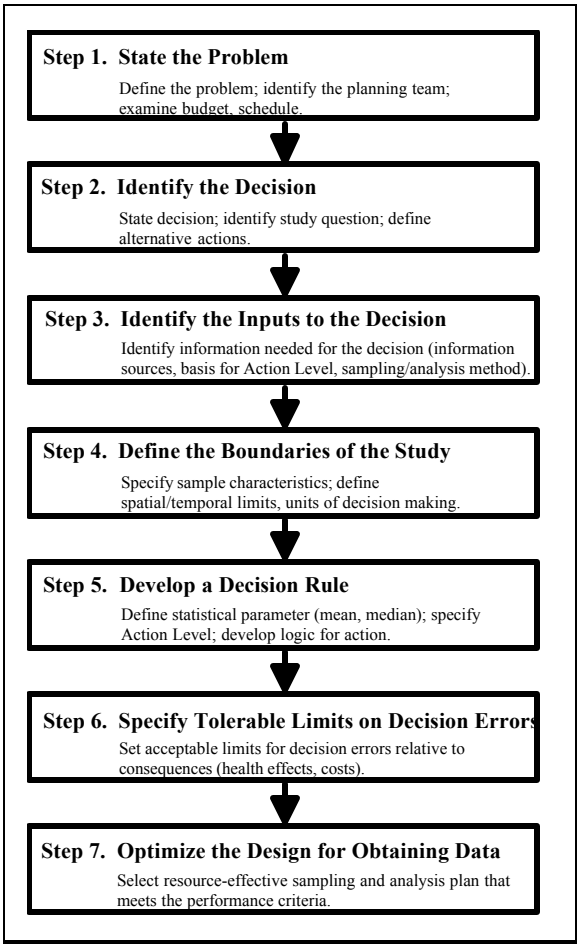


**Figure 3-1. The DQO Process**

define the population of interest, defines the spatial and temporal boundaries, defines the scale of decision making, and identifies any practical constraints on data collection.

**Step 5: Develop a Decision Rule.** This step develops a decision rule, a statement that allows the decision maker a logical basis for choosing among alternative actions, by determining the parameter of interest, action level, scale of decision making, and outlining alternative actions.

**Step 6: Specify Tolerable Limits on Decision Errors.** This step determines the decision maker's tolerable limits on potential decision errors by identifying the decision errors and base-level assumptions, specifying a range of possible parameter values where the consequences of decision errors are relatively minor, and assigning probability values to the probability for the occurrence of potential decision errors.

**Step 7: Optimize the Design for Obtaining Data.** This final step identifies a resource-effective sampling design for data collection for generating data. This design is then expected to satisfy the DQOs. Meeting or exceeding the DQOs is the goal of selection of sampling design.

By using the DQO Process, the planning team clarifies study objectives, defines the appropriate types of data, and specifies tolerable levels of potential decision errors that will be used to establish the quality and quantity of data needed to support decisions. Through this process, the planning team can examine trade-offs between the uncertainty of results and cost of sampling and analysis in order to develop designs that are acceptable to all parties involved. These are all important inputs to the sampling design process.

**What information will be needed to implement the sampling design process?**

The information needed includes outputs from the systematic planning process (for example, the outputs from Steps 1 through 6 of the DQO Process) and specific information about contributing factors about the specific problem that could influence the choice of design. The categories of factors that should be used in developing a sampling design are shown in Figure 3-2 and include:

Information About the Process or Area of Concern includes the conceptual model and any additional information about the process or area (for example, any secondary data from the site that are available, including results from any pilot studies).

Data Quality Information that is needed as input to the sampling design process is mainly from the DQO Process and include:

- The purpose of the data collection—that is, hypothesis testing (evidence to reject or support a finding that a specific parameter exceeds a threshold level, or evidence to

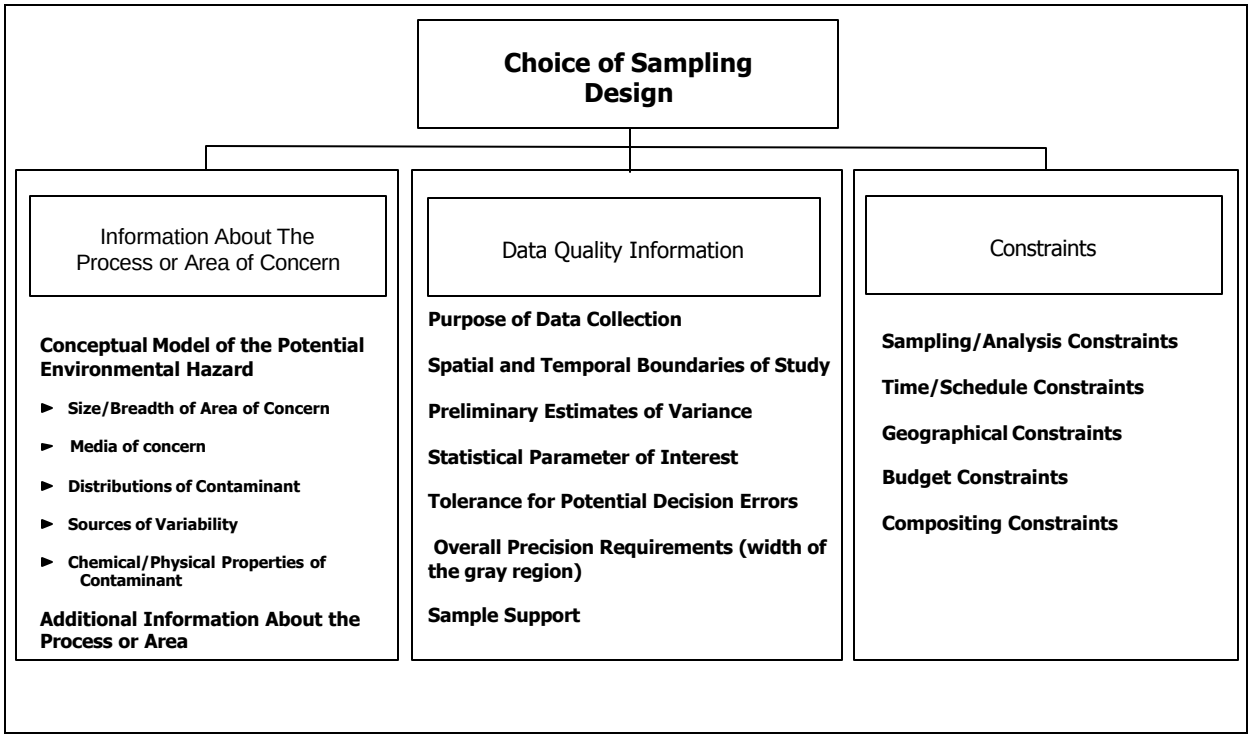


**Figure 3-2. Factors in Selecting a Sampling Design**

reject or support a finding that the specified parameters of two populations differ), estimating a parameter with a level of confidence, or detecting hot spots (DQO Step 5).

- The target population and spatial/temporal boundaries of the study (DQO Step 4).

- Preliminary estimation of variance (DQO Step 4).

- The statistical parameter of interest, such as mean, median, percentile, trend, slope, or percentage (DQO Step 5).

- Limits on decision errors and precision, in the form of false acceptance and false rejection error rates and the definition of the gray region (overall precision specifications) (DQO Step 6).

Constraints are principally sampling design and budget.

For more details on the DQO Process see *Guidance on the Data Quality Objectives Process (QA/G-4)* (EPA, 2000b).

It is important to carefully consider early in the design phase the sample support of the data to be collected and the proposed method of conducting the chemical analysis. The sample support is the physical size, shape, and orientation of material that is extracted from the sampling unit and subjected to the measurement protocol. In other words, the sample support comprises the portion of the sampling unit that is actually available to be measured or observed, and therefore to represent the sampling unit. Consequently, the sample support should be chosen so that the measurement protocol captures the desired characteristics of the sampling unit, given the inherent qualities of and variability within the sampling unit, and is consistent with the objectives of the study. The specification of sample support also should be coordinated with the actual physical specifications of the chosen analytical method(s) to ensure that a sufficient quantity of material is available to support the needed analyses. Usually, the analytical method needs a much smaller amount of material than that needed for the sample support to represent the sampling unit. In that case, the measurement protocol will specify how the sample support will be processed and subsampled to yield the amount of material needed for analysis.

Some examples will help clarify how sample support relates to sampling units and analytical methods. Consider a study that is designed to estimate average arsenic contamination in surface soil at a site. The project team may decide to divide the site into square sampling units that are 3 meters on each side and 10 centimeters deep. Given their knowledge of variability experienced at other sites, the project team may decide that the sample support needed to properly characterize a sampling unit is the area and volume of soil that can be obtained by taking 9 soil cores, each 15 cm in diameter and 10 cm deep. Consider another example in which a study is designed to estimate average mercury contamination in fish. The project team may decide that the sampling unit is an individual fish, and the sample support is the type and mass of fish tissue extracted from each fish, which they might specify in a table. In both of the above examples, an analytical chemist would confirm that the sample support would provide a sufficient amount of soil or fish tissue to conduct the analytical procedures needed to characterize the concentrations of arsenic in soil or mercury in fish. Sometimes the sample support is an integral part of the analytical result. For example, when sampling water for the occurrence of microbiological contaminants such as *chryptosporidium*, water is passed through filters and the filters are then processed and examined to count the number of organisms. The volume of water filtered constitutes the sample support and also is used directly in the calculation of the occurrence rate (i.e., number of organisims per volume of water). In all cases, the sample support is chosen to ensure that the measurement protocol will reliably characterize the sampling unit in a way that is consistent with the study objectives. The study objectives are defined during systematic planning, such as in DQO Steps 1 and 2. The definition of the sampling unit and selection of sampling support will depend strongly on the study boundaries defined in DQO Step 4, and on the performance criteria developed in DQO Step 6.

Possible constraints on choosing a sampling design fall into four categories: sampling/analysis limitations, time/schedule restrictions, geographic barriers, and budget amounts. Sampling/analysis constraints could include measurement instrument performance (for example, sensitivity and selectivity requirements for field or laboratory technologies), regulatory requirements that specify analytic or

sample collection method, or weather constraints (for example, performance of field technologies at low temperature, high humidity, or the ability to collect samples during certain seasons or types of weather). Time/schedule constraints could include seasonal constraints such as the relationship of exposure to season (for example, solvent volatility in warmer weather) and the availability of certified professionals. Geographic constraints could include physical barriers that may preclude sampling (for example, rivers, fences) and also any possible hindrance to the ability to accurately identify sample location. Budget constraints should take into account the entire data collection process—from the collection of the sample in the field, including transport and storage, to analysis of the samples and data entry and validation. Compositing constraints could include the decision on representativeness of the physical sample taken at a location or station, or the ability to physically mix samples both in the field and in the laboratory.

In addition to these categories, sampling design development should also take into account existing regulations and requirements (for example, state, municipal) if they apply. Finally, any possible secondary uses of the data should be considered to the extent possible.

## 3.3 STEPS IN THE SAMPLING DESIGN PROCESS

Steps of the sampling design process are represented in Figure 3-3 and described below.

**Review the systematic planning outputs**. First, the sampling objectives need to be stated clearly. Next, make sure the acceptance or performance criteria are specified adequately (such as probability limits on decision errors or estimation intervals). Then review the constraints regarding schedule, funding, special equipment and facilities, and human resources.

**Develop general sampling design alternatives.** Decide whether the approach will involve episodic sampling events (where a sampling design is established and all data for that phase are collected according to that design) or an adaptive strategy (where a sampling protocol is established and sampling units are selected in the field, in accordance with the protocol, based on results from previous sampling for that phase). Consider sampling designs that are compatible with the sampling objectives. Evaluate advantages, disadvantages, and trade-offs in the context of the

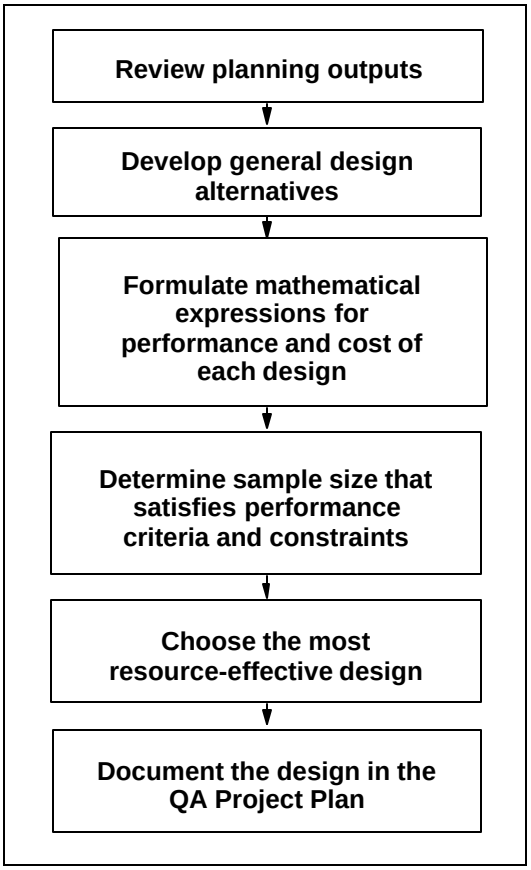


**Figure 3-3. The Sampling Design Process**

specific conditions of the study including the anticipated costs for possible alternative sampling strategies.

**Formulate mathematical expressions for the performance and cost of each design alternative.** For each design, develop the necessary statistical model or mathematical formulae needed to determine the performance of the design, in terms of the desired statistical power or width of the confidence interval. This process usually involves developing a model of relevant components of variance and estimating the total variance, plus key components as necessary. Also for each design, develop a cost model that addresses fixed costs (such as mobilization and setup costs) and variable costs (such as labor hours per sample and analytical costs per sample). Note that this step is not used in judgmental sampling designs. Assistance from a statistician will be needed to develop these formulae for more complex designs; formulae for the simpler designs are provided in the appendices to the chapters in this guidance.

**Determine the sample size that satisfies the performance criteria and constraints.** Calculate the optimal sample size (and sample allocation, for stratified designs or other more complex designs). This guidance document provides formulae for estimating sample sizes needed for the different designs. Trade-offs may be needed between less precise, less expensive measurement protocols (that allow for more sampling units to be selected and measured) and more precise, more expensive measurement protocols (that provide better characterization of each sampling unit at the expense of allowing fewer sampling units to be selected and measured). Care has to be taken to ensure that the trade-offs made do not change the inferences from the initially planned design. For example, the use of compositing designs needs to agree with the initial concepts of exposure or goal of the study.

If none of the designs are feasible (i.e., performance specifications cannot be satisfied within all constraints), then consider the following possible corrective actions listed below. Note that this step is not used in judgmental sampling designs because performance criteria are not explicitly considered.

- Consider other, more sophisticated, sampling designs.
- Relax performance specifications (for example, increase the allowable probability of committing a decision error) at the expense of increasing decision error risk.
- Relax one or more constraints (for example, increase the budget).
- Reevaluate the sampling objectives (for example, increase the scale of decision making, reduce the number of sub-populations that need separate estimates, or consider surrogate or indicator measurements).

**Choose the most resource-effective design.** Consider the advantages, disadvantages, and trade-offs between performance and cost among designs that satisfy performance specifications and constraints. Consider practical issues, schedule and budget risks, health and safety risks to project

personnel and the community, and any other relevant issues of concern to those involved with the project. Finally, obtain agreement within the planning team on the appropriate design.

**Document the design in the QA Project Plan.** Provide details on how the design should be implemented, contingency plans if unexpected conditions or events arise in the field, and quality assurance (QA) and quality control (QC) that will be performed to detect and correct problems and ensure defensible results. Specify the key assumptions underlying the sampling design, particularly those that should be verified during implementation and assessment. Details on how to write a QA Project Plan can be found in *Guidance for Quality Assurance Project Plans (QA/G-5)* (EPA, 1998b).

## 3.4 SELECTING A SAMPLING DESIGN

Table 3-1 presents examples of problem types that one may encounter and suggests sampling designs that are relevant for these problem types in particular situations.

**Table 3-1. Choosing the Appropriate Sampling Design for Your Problem**

| If you are... | and you have... | consider using... | in order to... |
|---|---|---|---|
| performing a screening phase of an investigation of a relatively small-scale problem | a limited budget and/or a limited schedule | judgmental sampling | assess whether further investigation is warranted that should include a statistical probabilistic sampling design. |
| developing an understanding of when contamination is present | an adequate budget for the number of samples needed | systematic sampling | acquire coverage of the time periods of interest. |
| developing an understanding of where contamination is present | an adequate budget for the number of samples needed | grid sampling | acquire coverage of the area of concern with a given level of confidence that you would have detected a hot spot of a given size. |
| estimating a population mean | an adequate budget | systematic or grid sampling | also produce information on spatial or temporal patterns. |
| | budget constraints and analytical costs that are high compared to sampling costs | composite sampling | produce an equally precise or a more precise estimate of the mean with fewer analyses and lower cost. |

**Table 3-1. Choosing the Appropriate Sampling Design for Your Problem**

| If you are... | and you have... | consider using... | in order to... |
|---|---|---|---|
| | budget constraints and professional knowledge or inexpensive screening measurements to assess the relative amounts of the contaminant at specific field sample locations | ranked set sampling | reduce the number of analyses needed for a given level of precision. |
| estimating a population mean or proportion | spatial or temporal information on contaminant patterns | stratified sampling | increase the precision of the estimate with the same number of samples, or achieve the same precision with fewer samples and lower cost. |
| delineating the boundaries of an area of contamination | a field screening method | adaptive cluster sampling | simultaneously use all observations in estimating the mean |
| estimating the prevalence of a rare trait | analytical costs that are high compared to sampling costs | random sampling and composite sampling | produce an equally precise (or a more precise) estimate of the prevalence with fewer analyses and lower cost. |
| attempting to identify population units that have a rare trait (for a finite population of units) | the ability to physically mix aliquots from the samples and then retest additional aliquots | composite sampling and retesting | classify all units at reduced cost by not analyzing every unit. |
| attempting to identify population unit(s) that have the highest contaminant levels (for a finite population of units) | the ability to physically mix aliquots from the samples and then retest additional aliquots | composite sampling and retesting | identify such units at reduced cost by not analyzing every unit. |

# CHAPTER 4

# JUDGMENTAL SAMPLING

## 4.1 OVERVIEW

Judgmental sampling refers to the selection of sample locations based on professional judgment alone, without any type of randomization. Judgmental sampling is useful when there is reliable historical and physical knowledge about a relatively small feature or condition. As discussed in *Quality Assurance Guidance for Conducting Brownfields Site Assessments* (EPA, 1998a), whether to employ a judgmental or statistical (probability-based) sampling design is the main sampling design decision. This design decision applies to many environmental investigations including Brownsfield investigations. An important distinction between the two types of designs is that statistical sampling designs are usually needed when the level of confidence needs to be quantified, and judgmental sampling designs are often needed to meet schedule and budgetary constraints.

Implementation of a judgmental sampling design should not be confused with the application of professional judgment (or the use of professional knowledge of the study site or process). Professional judgment should *always* be used to develop an efficient sampling design, whether that design is judgmental or probability-based. In particular, when stratifying a population or site, exercising good professional judgment is essential so that the sampling design established for each stratum is efficient and meaningful.

## 4.2 APPLICATION

For soil contamination investigations, judgmental sampling is appropriate for situations in which any of the following apply:

- Relatively small-scale features or conditions are under investigation.
- An extremely small number of samples will be selected for analysis/characterization.
- There is reliable historical and physical knowledge about the feature or condition under investigation.
- The objective of the investigation is to screen an area(s) for the presence or absence of contamination at levels of concern, such as risk-based screening levels (note that if such contamination is found, follow-up sampling is likely to involve one or more statistical designs).
- Schedule or emergency considerations preclude the possibility of implementing a statistical design.

Judgmental sampling is sometimes appropriate when addressing site-specific groundwater contamination issues. As further discussed in *Quality Assurance Guidance for Conducting Brownfields Site Assessments* (EPA, 1998a), a statistical sampling design may be impractical if data are needed to evaluate whether groundwater beneath a Brownfields site is contaminated due to the high cost of groundwater sample collection and knowledge of the connection between soil and groundwater contamination.

## 4.3 BENEFITS

Because judgmental sampling designs often can be quickly implemented at a relatively low cost, the primary benefits of judgmental sampling are to meet schedule and budgetary constraints that cannot be met by implementing a statistical design. In many situations, when some or all of the conditions listed in Section 4.2 exist, judgmental sampling offers an additional important benefit of providing an appropriate level of effort for meeting investigation objectives without excessive consumption of project resources.

## 4.4 LIMITATIONS

Judgmental sampling does not allow the level of confidence (uncertainty) of the investigation to be accurately quantified. In addition, judgmental sampling limits the statistical inferences that can be made to the units actually analyzed, and extrapolation from those units to the overall population from which the units were collected is subject to unknown selection bias.

## 4.5 IMPLEMENTATION

By definition, judgmental sampling is implemented in a manner decided by the professional(s) establishing the sampling design. Specialized academic and professional training is needed before a professional is qualified to design a judgmental sampling program. The following paragraphs provide only a few examples of the most common factors that professionals should consider when establishing judgmental sampling designs.

As discussed in EPA's *Soil Screening Guidance* (EPA, 1996a), current investigative techniques and statistical methods cannot accurately establish the mean concentration of subsurface soils within a contaminated source without a costly and intensive sampling program that is well beyond the level of effort generally appropriate for screening. The *Soil Screening Guidance* advises that, in establishing a judgmental sampling design to investigate subsurface soil contamination, the professional should locate two or three soil borings in the areas suspected of having the highest contaminant concentrations. If the mean contaminant concentration calculated for any individual boring exceeds the applicable numerical screening value, additional investigative phases should be conducted. The *Soil*

*Screening Guidance* provides several approaches for calculating a mean contaminant concentration for each boring; these approaches vary with the sampling-interval design.

In establishing a judgmental sampling design to investigate a subsurface soil contamination problem, the professional needs to consider many factors including the following:

- Soil properties that affect contaminant migration (for example, texture, layering, moisture content);
- The physical and chemical nature of the contaminant under investigation (for example, solubility, volatility, reactivity);
- The manner in which the contaminant is understood to have been released (for example, surface spill, leachate generated through above ground or buried waste, leaking underground tank or pipe);
- The timing and duration of the release; and
- The amount of contaminant understood to have been released.

As stated in Section 4.2, judgmental sampling is often appropriate when addressing site-specific groundwater contamination issues. The most common factors to consider in establishing a judgmental sampling design to address a site-specific groundwater contamination issue include the following:

- The physical and chemical nature of the contaminant under investigation (for example, solubility, volatility, reactivity, density [whether floating or sinking nonaqueous phase liquid could be present]);
- The possible effects of contaminant migration through the unsaturated zone when and where the contaminant entered the aquifer;
- The possible ways that contaminant migration through the unsaturated zone might have changed the chemical nature of the contaminant before it entered the aquifer;
- The depths and thicknesses of aquifers beneath the site;
- The direction and rate of groundwater flow within each aquifer and variations in these parameters;
- The aquifer properties that cause the contaminant to disperse within it, both laterally and vertically; and
- The natural attenuation processes that may affect how the contaminant migrates in groundwater.

## 4.6 RELATIONSHIP TO OTHER SAMPLING DESIGNS

Other sampling designs are used in conjunction with judgmental sampling in two common situations. First, they may be used when the population or site is stratified, and judgmental sampling takes place within one or more strata. This situation is typical of small-scale soil contamination

investigations when the suspected location of the contaminant release is known. When the suspect area is identified as a stratum, then a judgmental sampling design is established for that stratum. Other strata established for the site may be addressed through implementation of statistical sampling designs. Judgment is, of course, used in establishing the boundaries and extent of each stratum.

Second, other sampling designs may be used when judgmental sampling indicates that the screening criteria established for the area under investigation is exceeded, thereby warranting further investigation. Depending on how much historical information is available and how much information has been obtained from the judgmental-sampling phase, follow-up phases of investigation might involve any of the statistical sampling designs described in this guidance document.

## 4.7 EXAMPLES OF SUCCESSFUL USE

### 4.7.1 Area Impacted by Contamination Can Be Visually Discerned

An active manufacturing facility is being sold, and the prospective purchaser is conducting an investigation to characterize existing environmental conditions and potential associated liability. One feature being assessed is an approximately 500 square meters ($m^2$) fenced area where drums of an aqueous cupric-chloride waste are stored. When released, the waste stains the soil blue-green. Eight irregularly shaped blue-green stains are identified ranging in size from about 10 square centimeters to a square meter. The stains are thought to be a result of relatively small releases that occurred as waste was poured into drums at the storage area from smaller containers filled at the facility's Satellite Accumulation Areas. A judgmental sampling design is established whereby a single grab sample of soil is collected from each of the observed stains and analyzed for copper concentration. If any single copper result falls within one order of magnitude of the risk-based copper soil-screening level for industrial land use, the seller has agreed to pay for a follow-up investigation that will involve a statistical sampling program designed to better characterize the soil copper contamination and assess whether remediation is warranted.

### 4.7.2 Potential Location of the Contaminant Release Is Known

An abandoned textile mill is being investigated as a Brownfields site, and one previous employee was located who gave a reliable account of site features and activities. Based on this interview, the site was stratified and several different sampling designs (some statistical and some judgmental) were established. A judgmental sampling design is being used to investigate a 30 meter long drain pipe that carried a variety of wastes from one of the site factories to a leach field adjacent to the building; a statistical grid-sampling design was established to investigate the leach field. The drain pipe is accessible under a grating installed on the basement floor of the factory, and visual (external) and video (internal) inspections of the pipe showed it to be in good condition with no observable deterioration or cracks. However, several of the joints between the 3 meter length pipe segments

appeared either loose or slightly separated. The judgmental sampling design established for this feature involved marking the basement floor adjacent to each pipe joint, removing the pipe, and collecting a single sample of the soil at each marked location for laboratory analysis. The analytical results then would be compared to the risk-based screening levels established for the list of potential site contaminants.

## 4.8 EXAMPLES OF UNSUCCESSFUL USE

### 4.8.1 Double Judgmental Sampling

Ginevan (2001) has a practical example:

> *"...a good question is 'what do I do if I am stuck with a "dirty spots" sample?' The answer is that if there is a great deal of money riding on the decision one should do the sampling over. Note also that nothing is ever so bad that it cannot be made worse. In one case we participated in, a dirty spots sample was taken first. This was pointed out to the client, who then went out and took a comparable number of samples from an area known to be clean. At this point the formula given by Land's procedure for the upper bound on the arithmetic mean of log-normal data was applied to the combined data (which were strongly bimodal because of the clean/dirty dichotomy). The resulting "upper bound" on the mean exceeded the largest observation from the dirty spots sample! Unhappily these data were beyond even the capability of the bootstrap to salvage. The original sample had been taken to find dirty spots and was thus not representative of the site. The end result was a set of about 100 measurements which told us almost nothing about the nature and extent of contamination at the site. The client then instituted a statistically designed sampling plan."*

### 4.8.2 Visual Judgmental Sampling

This example concerns a rural county enforcement officer tramping along a creek periodically exclaiming, "Here is a contamination!" when encountering dark spots in the stream sediment. Obviously, the samples collected were only representative of those "dark" areas of sediment declared contaminated by the enforcement officer and resulted in a wide range on concentration. Subsequent investigation of the support of color blind grab samples of sediment revealed that the variation within an areal area the size of a desk top encompassed all concentrations from not detected to those measured

by the enforcement officer. The support of the sample collected by the enforcement officer was no better than a single random grab sample.

These examples show how it is possible to be completely misled by reliance on what seems to be a desirable characteristic upon which to base the inclusion of a sample unit into the overall sample. The advantage gained by using a probabilistic sampling scheme is that such biases are avoided.

# CHAPTER 5

# SIMPLE RANDOM SAMPLING

## 5.1 OVERVIEW

Simple random sampling is the simplest and most fundamental probability-based sampling design. Most of the commonly used statistical analysis methods assume either implicitly or explicitly that the data were obtained using a simple random sampling design.

A simple random sample of size *n* is defined as a set of *n* sampling units selected from a population (of objects or locations in space and/or time) so that all possible sets of *n* sampling units have the same chance of being selected. For example, if there is a population of four elements (A,B,C,D) and a sample of size n=3 elements is drawn, without replacement, there are four possible outcomes:

(A,B,C), (A,B,D), (A,C,D), and (B,C,D).

Any sampling design that makes these outcomes equally likely is, by definition, a simple random sampling design. A simple random sample of size *n* occurs when *n* units are independently selected at random from the population of interest.

The most important characteristic of simple random sampling is that it protects against the bias (systematic deviation from the "truth") that can occur if units are selected subjectively. Because it is the most fundamental sampling design, simple random sampling also is a benchmark against which the efficiency and cost of other sampling designs often are compared. Moreover, when using an alternative sampling design, the minimum sample size (number of sampling units) needed for that sampling design often is estimated by first computing the sample size that would be needed with a simple random sampling design. That sample size is then multiplied by an adjustment factor, called the survey design effect, to produce the minimum sample size needed under the alternative sampling design [Section 4.1.1 of Cochran (1977)].

## 5.2 APPLICATION

Simple random sampling is appropriate when the population being sampled is relatively uniform or homogeneous. In practice, simple random sampling usually is used in conjunction with other sampling designs, as discussed in Section 5.6.

Simple random sampling often is appropriate for the last stage of sampling when the sampling design has more than one stage of sampling (i.e., a sample of units is selected at the first stage and then

subunits are selected from each sample unit) [Chapter 6 of Gilbert (1987) and Chapters 12 and 13 of Thompson (1992)]. Examples include the following:

- Selecting one or more leaves from each sample plant for characterization,
- Selecting one or more aliquots from each soil sample for chemical analysis, and
- Assigning split samples or aliquots to laboratories or analytical methods.

In a similar vein, simple random sampling usually is needed for assigning experimental units to treatments, or experimental conditions, in experimental designs.

## 5.3 BENEFITS

The primary benefit of simple random sampling is that it protects against selection bias by guaranteeing selection of a sample that is representative of the sampling frame, provided that the sample size is not extremely small (for example, 20 observations or more). Moreover, the procedures needed to select a simple random sample are relatively simple.

Other benefits of using simple random sampling include the following:

- Statistical analysis of the data is relatively straightforward because most common statistical analysis procedures assume that the data were obtained using a simple random sampling design.

- Explicit formulae, as well as tables and charts in reference books, are available for estimating the minimum sample size needed to support many statistical analyses.

## 5.4 LIMITATIONS

Simple random sampling has two primary limitations:

- Because all possible samples are equally likely to be selected, by definition, the sample points could, by random chance, not be uniformly dispersed in space and/or time. This limitation is overcome somewhat as the sample size increases, but it remains a consideration, even with a large number of samples.

- Simple random sampling designs ignore all prior information, or professional knowledge, regarding the site or process being sampled, except for the expected variability of the site or process measurements. Prior information almost always can be used to develop a probability-based sampling design that is more efficient than simple random sampling (i.e., needs fewer observations to achieve a given level of precision).

Because of these limitations, simple random sampling is seldom recommended for use in practice except for relatively uniform populations. Stratified simple random sampling (Chapter 6) is commonly used to overcome these limitations by defining geographic and/or temporal sampling strata. Alternatively, one may use systematic sampling (Chapter 7) or quasi-random sampling (Section 5.5.2) to overcome these same limitations. Nevertheless, simple random sampling is a fundamental building block and benchmark for most other sampling designs.

## 5.5 IMPLEMENTATION

This section discusses how to determine the minimum sample size needed with simple random sampling to (1) estimate a population mean or proportion with prespecified precision or (2) test a hypothesis regarding a population mean or proportion with a prespecified significance level and power. This section also addresses the process of selecting a simple random sample.

### 5.5.1 How do you estimate the sample size?

To determine the minimum sample size needed to estimate a population proportion (for example, proportion of units with concentrations above a health-based threshold), first identify a conservative preliminary estimate of the true population proportion. In the absence of prior information, use 50% as the preliminary estimate as this results in the largest sample size and so is the most conservative. The closer the preliminary estimate is to the actual value, the greater the savings in resources.

To determine the minimum sample size needed to estimate a population mean (for example, mean contaminant concentration), first identify a conservatively large preliminary estimate of the population variance. The preliminary estimate should be large enough that the true population variance is not likely to be larger than the preliminary estimate because the sample size will be too small if the estimated variance is too small. Sources of a preliminary estimate of population variance include: a pilot study of the same population, another study conducted with a similar population, or an estimate based on a variance model combined with separate estimates for the individual variance components. In the absence of prior information, estimate the standard deviation (square root of the variance) by dividing the expected range of the population by six, i.e.

$$\hat{\sigma} = \frac{\text{Expected Maximum} - \text{Expected Minimum}}{6}$$

However, this is only a crude approximation and should be used only as a last resort.

Using these inputs, Appendix 5 provides general-purpose formulae for determining the minimum sample size needed to achieve specified precision for estimates of population means and

proportions. Sample size formulae for achieving specified power for hypothesis tests are in Section 3 of *Guidance for Data Quality Assessment (QA/G-9)* (EPA, 2000a). Appendix 5 tabulates the results from applying these formulae for determining the minimum sample size needed for hypothesis tests. Examples of the use of these tables are provided in Section 5.7.2.

If the sample sizes calculated using the simple random sampling formulae are greater than the study budget can support, then other sampling designs may reduce the number of sample specimens and/or the number of measurements. For example, stratified random sampling (Chapter 6) and ranked set sampling (Chapter 8) may result in smaller sample sizes if (inexpensive) data are available that are positively correlated with the outcomes of interest. Moreover, if the objective of the study is estimation of means, composite sampling (Chapter 10) may greatly reduce the number of analytical measurements. Finally, if the variability between replicate measurements (for example, in the lab) is greater than the natural variability between units (for example, using an imprecise method to analyze water samples from a fairly homogenous body of water), using the mean of replicate measurements on each sample specimen may reduce the number of sample specimens.

### 5.5.2 How do you decide where to take samples?

Selecting a simple random sample is most straightforward when all the sampling units (for example, barrels in a warehouse, trees at a study site) comprising the population of interest can be listed. When selecting a simple random sample from a list of $N$ distinct sampling units, use the following procedure:

- Label the sampling units from *1* to $N$.
- Use a table of random numbers, or a computerized random number generator, to randomly select $n$ integers from *1* to $N$ from the list.

The set of sampling units with these $n$ labels comprises a simple random sample of size $n$. These $n$ sample units may be $n$ points on the surface of a hazardous waste site, $n$ points in time, etc. Here the word "sample" is used in this statistical sense, related to a list of sampling units or potential sampling locations. The actual aliquots of air, water, soil, etc., that are collected at the sample locations are referred to as sample "specimens" to distinguish them from the statistical sample selected from the universe of all possible sampling units (objects or locations in space and/or time).

When *selecting a sample from a two-dimensional medium*, such as surface soils or the bottom of a lake or stream, the above one-dimensional list sampling approach can be used if an $M$ by $N$ grid is used to partition the population into $MN$ unique units and the sample is selected from the list of $MN$ units.

However, it is often more practical and flexible to select points directly at random in two-dimensional space if the desired sample support is not a rectangular area. If a rectangular coordinate system (i.e., $x$ and $y$ coordinates, such as latitude and longitude) can be superimposed on the area of interest, then a simple random sample of points is generated by randomly generating $x$- and $y$-coordinates, as illustrated in Figure 5-1. Note that in an irregularly shaped sample area, randomly generated points falling outside of the sample area are not used.

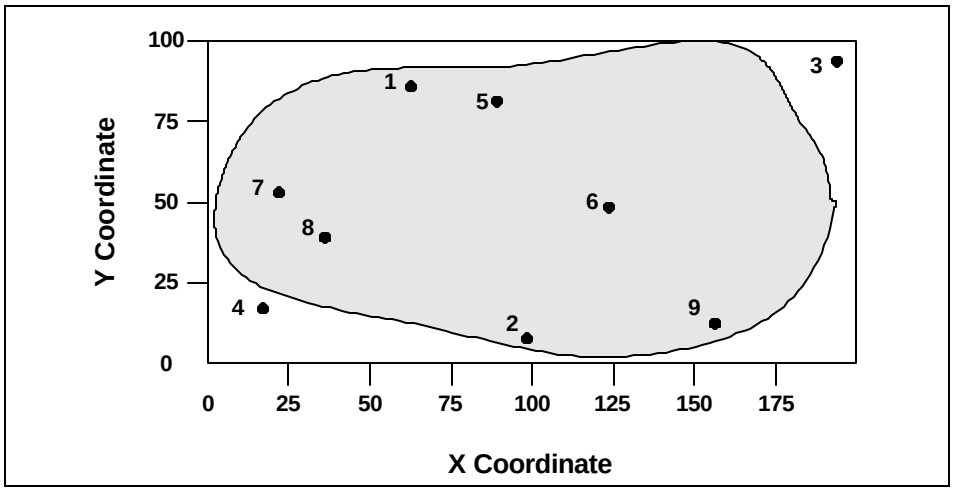


**Figure 5-1. Example of a Map Showing Random Sampling Locations**

When these sampling procedures are implemented to generate simple random samples in two dimensions, the randomly generated sampling points (i.e., $x$- and $y$- coordinates or direction) should be rounded to the nearest unit that can be reliably identified in the field (for example, nearest 1 or 5 meters). A sample specimen with the support defined in the sampling plan should then be obtained as near as possible to each of these approximate random sampling points using a procedure to avoid subjective bias factors such as "difficulty in collecting a sample, the presence of vegetation, or the color of the soil" (EPA, 2000b). The protocols should be defined so that it will always be possible to obtain a sample from each randomly selected location. However, if it is physically impossible to obtain a specimen from a randomly selected location, deleting that location from the sample is valid as long as inferences are restricted to the accessible locations. The use of a subsidiary list of alternate (random) locations to be substituted for inaccessible locations is recommended.

The above sampling methods can be extended fairly easily, at least conceptually, to *sampling three-dimensional wastes* (for example, a waste pile or liquid wastes in a pond, lagoon, or drum). One approach is to superimpose a three-dimensional coordinate system over the area to be sampled (i.e., $x$, $y$, and $z$ coordinates) and randomly generate $x$-, $y$-, and $z$-coordinates to identify randomly selected points.

Although it is conceptually easy to generate random sampling points in three dimensions, actually getting a sampling tool into a three-dimensional medium at these randomly selected locations and extracting specimens with the correct sample support (size, shape, and orientation) can be difficult or impossible. Consider, for example, solid waste in a pile. If the waste pile has the consistency of soil, a technician may be able to take a core sample at the randomly selected location and extract a subsample from the core at the correct depth that has the desired support (for example, 5 centimeter diameter and 15 centimeters depth). However, if the pile contains large impermeable solids (for example, rocks of larger diameter than the core), taking such a core sample may not be possible. Alternatively, if the material is very fine, like ash, a technician may not be able to take a core sample because the process of getting the core would fundamentally alter the nature of the pile being sampled (for example, it would cause the pile to shift or collapse). In that case, one potential solution may be to level the pile and take samples from the entire depth of the leveled pile at randomly selected points in two dimensions.

Liquid wastes present similar problems for sampling in three dimensions. If the liquid waste has the consistency of water, it may be possible to extract samples from randomly selected locations using a probe and pump. However, some wastes (for example, a semiliquid sludge) are too thick to be pumped yet not solid enough to extract competent cores. If a technician were sampling sludge from a lagoon, it might be necessary to sample the entire vertical thickness of sludge at randomly selected locations (in two dimensions) and then analyze a subsample(s) from the resulting composite sample.

Section 21.6.5 of Pitard (1993) states that one could theoretically obtain correct (representative) samples from a waste pile by selecting either one- or two-dimensional samples representing the full cross-section of the waste. A one-dimensional sample is one in which vertical cross-sections of a prescribed thickness are selected, as depicted in Figure 5-2. A two-dimensional sample is one in which cores from the top to the bottom of the waste pile are randomly extracted, as depicted in Figure 5-3. Section 14.4.7 of Pitard (1993) states that attempting to extract such samples is an "exercise in futility" because of the lack of appropriate sampling devices. Additional guidance regarding sampling

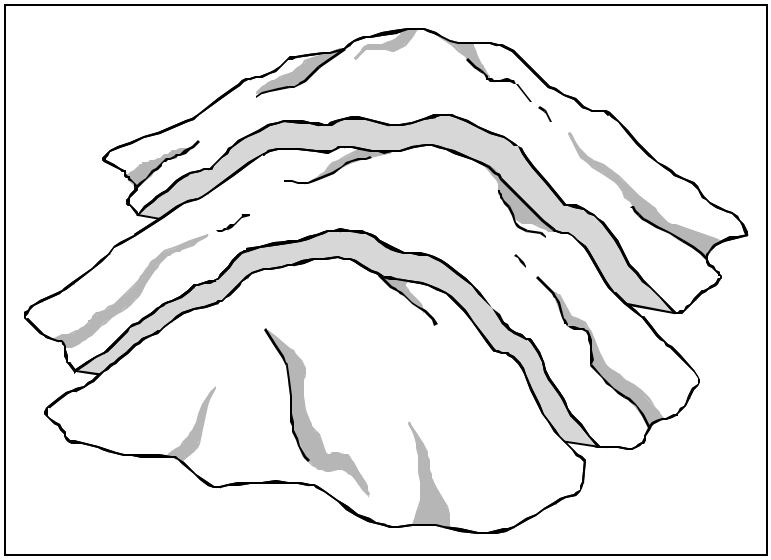

**Figure 5-2. A One-Dimensional Sample of Cross-Sections from a Waste Pile**

devices and techniques that can be used to sample from three-dimensional waste piles is provided in Section 8.3 of Myers (1997) and by the American Society for Testing and Materials (ASTM) D6232-00 (2000).

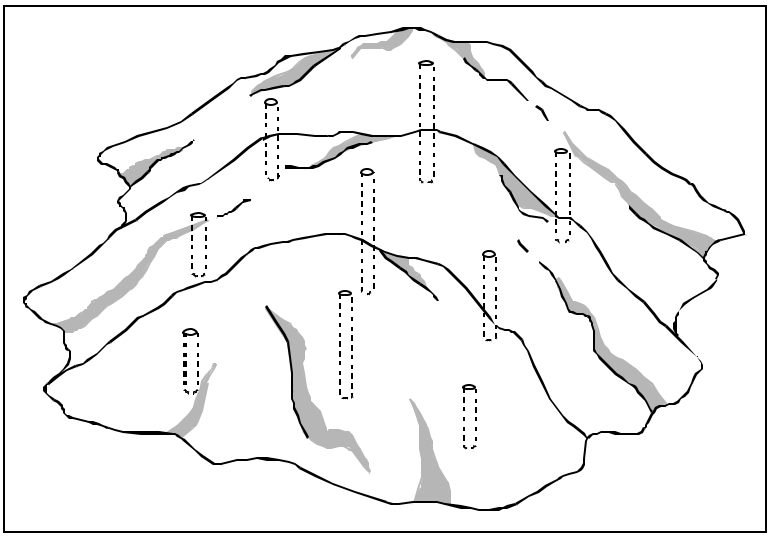

**Figure 5-3. A Two-Dimensional Sample of Cores from a Waste Pile**

An alternative sampling method that provides random samples that are more uniformly dispersed than simple random samples is "quasi-random sampling." Quasi-random sampling refers to methods for generating a quasi-random sequence of numbers that are "in a precise sense, 'maximally avoiding' of each other" [Section 7.7 of Press et al.(1992)]. Samples in two or more dimensions are generated by pairing two or more of these quasi-random sequences. In two dimensions, the result is a set of sample points that, for any given sample size, appear to be uniformly scattered throughout the sampled area, as illustrated in Figure 5-4. Quasi-random sampling can be used to avoid the potential for geographic clustering that exists with simple random sampling without taking the risk of aligning the sample with an unknown pattern of contamination, a limitation of grid sampling (as discussed in Chapter 7). The resulting data can be analyzed as if the sample were a simple random sample, knowing that the sampling variance is likely to be slightly underestimated. Techniques for generating quasi-random samples are mathematically complex; they are described in Section 7.7 of Press et al. (1992). A simpler technique that achieves similar results is "deep" stratification, in which only one unit is selected at random from each sampling stratum (see Chapter 6). A variation would be to divide the population into small units and take a random sample from within each unit for a total of *n* units.

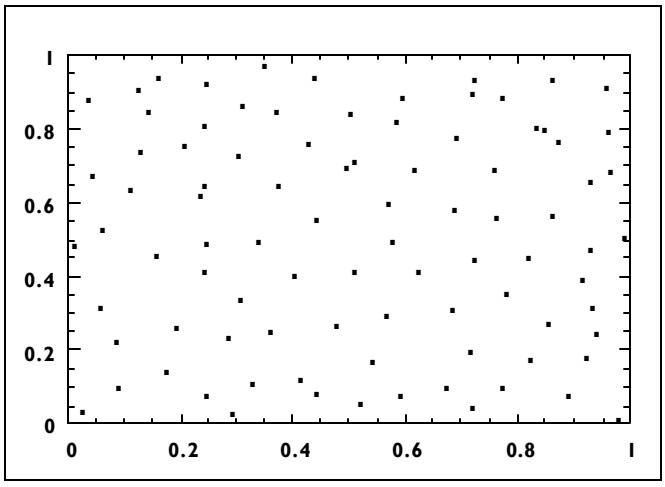


**Figure 5-4. Illustration of a Quasi-Random Sample**

## 5.6 RELATIONSHIP TO OTHER SAMPLING DESIGNS

Simple random sampling often is used for selecting samples within sampling strata. When an independent simple random sample is selected from each stratum, the sampling

design is referred to as stratified simple random sampling (see Chapter 6). Simple random sampling also is used as the first step of the ranked set sampling process described in Chapter 7. It also can be used as the first step of the adaptive cluster sampling process described in Chapter 9.

## 5.7 EXAMPLES

### 5.7.1 General Simple Random Sampling Example

Suppose that a company with a fleet of 5,000 late-model, mid-sized sedans decides that they will overhaul their fleet to improve emissions if the mean (average) carbon monoxide (CO) emission rate of the fleet (in grams per mile, g/m) is unusually high. Since the EPA standard for passenger cars is no more than 3.4 g/m, and data from the manufacturers of their fleet's cars suggests that most cars in the fleet will be between 1.0 and 3.0 g/m, they decide that an overhaul is needed if their mean CO emission rate exceeds 2.5 g/m. Hence, to determine whether or not an overhaul is needed, they will test the following hypothesis for means:

$$H_0: \ : \# 2.5 \text{ versus } H_A: \ : > 2.5 \text{ g/m}$$

Suppose that all vehicles in the fleet are late-model, 6-cylinder cars that are expected to have similar emission rates. Hence, for selecting a sample of vehicles to be tested from this relatively homogeneous 5,000-vehicle population, a simple random sampling design is appropriate.

In order to determine appropriate sample sizes using Appendix Table 5-1, a preliminary estimate of the variability between measurements of CO emission rates is needed for their fleet. Company researchers referred to old records to estimate the expected variability in the fleet's CO emission rates. However, lacking any data regarding variances of CO emission rates, they choose to use one-sixth of the expected range as an estimate of the standard deviation, as discussed in Section 5.5.1. They expected that the range probably would be from about 0.5 to 3.5, a range of 3.0 g/m. and were could potentially be as large as 4 g/m or more if some of their cars were not properly tuned. Hence, sample sizes were determined for the following potential standard deviations:

| **Range** (g/m) | $\hat{s} = Range / 6$ |
|---|---|
| 3 | 0.50 |
| 4 | 0.67 |
| 5 | 0.83 |

In their application of the DQO Process, the company officials determined that the maximum acceptable error rates were as follows:

- False Rejection: " = Prob(false rejection when : = 2.5 g/m) = 0.05
- False Acceptance: $ = Prob(false acceptance when : = 2.75 g/m) = 0.05

Table 5-1 then was used to determine the minimum sample size needed by entering the table with the following parameters:

- " = Significance level = 0.05 (i.e. 5%)
- Power = 1 - $ = 0.95 (i.e. 95%)
- Effect size 1 = 100( $|:_1 - :_0|$ /F) = 100( |2.75 - 2.50| ) / 0.50 = 50%
- Effect size 2 = 100( $|:_1 - :_0|$ /F) = 100( |2.75 - 2.50| ) / 0.67 = 37%
- Effect size 3 = 100( $|:_1 - :_0|$ /F) = 100( |2.75 - 2.50| ) / 0.83 = 30%

Hence, the company managers used the first row of Table 5-1 to determine that a sample of 122, 69, or 45 cars was needed, depending on whether the effect size was 30%, 40%, or 50%, respectively. Based on these results, they decided that a simple random sample of 100 cars should provide adequate protection against both false rejection and false acceptance decision errors.

The researchers then assigned inventory control numbers to the cars in the fleet from 1 to 5,000 to facilitate the random sampling process. They used a random number generator to generate 100 random numbers between 1 and 5,000 (for example, using http://www.random.org). The cars with these inventory control numbers were then selected as the simple random sample of cars to be tested for CO emission rates.

In this case, the cost of sampling (measuring the emission rate) was relatively low and a large sample presented no problems. If the cost had been prohibitive, a pilot study would have been completed in order to give preliminary information on the variability. This would probably result in a lower number of cars to test.

### 5.7.2 Examples Using Look-up Tables in Appendix 5

These examples are simply intended to demonstrate the use of the tables.

**Tables 5-2 and 5-3:** Suppose the company decides that they need to overhaul the fleet of cars if more than 10% of the fleet have CO emission rates exceeding 3.0 g/m. To determine whether or not the overhaul is needed, they need to test the hypothesis for proportions:

$$H_0: P \# 10\% \text{ versus } H_A: P > 10\%$$

In their application of the DQO Process, the company officials determine that the maximum acceptable error rates are as follows:

- False Rejection: $\alpha$ = Prob(false rejection when P = 10%) = 0.05
- False Acceptance: $\beta$ = Prob(false acceptance when P = 15%) = 0.05

Table 5-2 then can be used to determine the minimum sample size needed by entering the table with the following parameters:

- $\alpha$ = Significance level = 0.05 (i.e., 5%)
- Power = 1 - $\beta$ = 0.95 (i.e., 95%)
- $P_0$ = 10%
- $|P_1 - P_0| = |15\% - 10\%| = 5\%$

Table 5-2 shows that a sample of 468 cars is necessary to achieve the error bounds specified for the hypothesis test.

**Table 5-4:** Suppose the company also has a fleet of 5,000 small pick-up trucks. The researchers want to know if the mean CO emission rate for their fleet of pick-up trucks exceeds that for the fleet of sedans. They then need to test the hypothesis for difference of two means:

$$H_0: \mu_1 - \mu_2 \leq 0 \text{ versus } H_A: \mu_1 - \mu_2 > 0,$$

where $\mu_1$ is the mean CO emission rate for the fleet of pick-up trucks and $\mu_2$ is the mean CO emission rate for the fleet of sedans.

In their application of the DQO Process, they determine that the maximum acceptable error rates are as follows:

- $\alpha$ = Prob(false rejection when $\delta = \mu_1 - \mu_2 = 0$) = 0.05
- $\beta$ = Prob(false acceptance when $\delta = \mu_1 - \mu_2 = 0.25$ g/m) = 0.05

Table 5-4 then can be used to determine the minimum sample size needed by entering the table with the following parameters:

- $\alpha$ = Significance level = 0.05 (i.e.5%)
- Power = 1 - $\beta$ = 0.95 (i.e.95%)
- Effect size = $100( |\delta_1 - \delta_0| / \sigma) = 100( |0.25 - 0.00| / 0.50) = 50\%$

Table 5-4 shows that a sample of 88 sedans and 88 pick-up trucks is necessary to achieve the error bounds specified for the hypothesis test.

**Tables 5-5 and 5-6:** Suppose the company decides that they want to determine whether the proportion of pickup trucks in the fleet with CO emission rates greater than 3.0 g/m is greater than the proportion for the fleet of sedans. They then need to test the hypothesis for difference of two proportions:

$$H_0: P_1 - P_2 \# 0\% \text{ versus } H_A: P_1 - P_2 > 0\%$$

where $P_1$ is the proportion of pick-up trucks with emission rates exceeding 3.0 g/m and $P_2$ is the proportion of sedans with emission rates exceeding 3.0 g/m.

In their application of the DQO Process, they determine that the maximum acceptable error rates are as follows:

- False Rejection: " = Prob(false rejection when $P_1 - P_2 = 0$) = 0.05
- False Acceptance: $ = Prob(false acceptance when $P_1 = 10\%$ and $P_2 = 5\%$) = 0.05

Table 5-5 then can be used to determine the sample size needed by entering the table with the following parameters:

- " = Significance level = 0.05 (i.e.5%)
- Power = 1 - $ = 0.95 (i.e.95%)
- $P_1 = 10\%$
- $|P_1 - P_{2|} = |10\% - 5\%| = 5\%$

Table 5-5 indicates that a sample of 947 sedans and a sample of 947 pick-up trucks are necessary to achieve the error bounds specified for the hypothesis test.

It should be noted, however, that when the estimated sample size (*n*) becomes relatively large compared to the population size (*N*), a factor called the Finite Population Correction Factor, the ratio *n/N*, must be taken into consideration. For more information, see Section 4.2 of Gilbert (1987), Section 2.5 of Cochran (1963), and Appendix 5. In addition, these formulae assume the underlying population to be normally distributed. If approximate normality does not hold, these sample sizes could be too small.

## APPENDIX 5

## SAMPLE SIZE TABLES FOR SIMPLE RANDOM SAMPLING DESIGNS

This appendix provides the following tables to determine the minimum sample size needed to achieve sufficient precision with simple random sampling designs:

- Table 5-1. Sample Size Needed for a One-Sample t-Test.
- Table 5-2. Sample Size Needed for a One-Sample Test for a Population Proportion, P, at a 5% Significance Level.
- Table 5-3. Sample Size Needed for a One-Sample Test for a Population Proportion, P, at a 10% Significance Level.
- Table 5-4. Sample Size Needed for a Two-Sample t-Test.
- Table 5-5. Sample Size Needed for a Two-Sample Test for Proportions at a 5% Significance Level.
- Table 5-6. Sample Size Needed for a Two Sample Test for Proportions at a 10% Significance Level.

The formulae that these sample size calculations are based upon are provided in Chapter 3 of *Guidance for Data Quality Assessment (QA/G-9)* (EPA, 2000a) for the remaining tables, which address sample size needed for hypothesis tests.

**Table 5-1. Sample Size Needed for One-Sample t-test**

| Significance Level | Power | Effect Size 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|
| 5% | 95% | 1,084 | 272 | 122 | 69 | 45 |
| | 90% | 858 | 216 | 97 | 55 | 36 |
| | 80% | 620 | 156 | 71 | 40 | 27 |
| 10% | 95% | 858 | 215 | 96 | 55 | 36 |
| | 90% | 658 | 166 | 74 | 42 | 28 |
| | 80% | 452 | 114 | 51 | 29 | 19 |

Case 1: $H_0$: : # *C* vs $H_A$: : > *C*; Case 2: $H_0$: : $ *C* vs $H_A$: : < *C*. In either case, the effect size is $100(|:_1 - .|)/F$, where $: = :_1$ is at the boundary of the gray region determined in Step 6 of the DQO Process and $F$ is a preliminary estimate of the population standard deviation (square root of the variance).

**Table 5-2. Sample Size Needed for a One-Sample Test for a Population Proportion, P, at a 5% Significance Level**

| $P_0$ | | $\lvert P_1 - P_0 \rvert$ | | | |
|---|---|---|---|---|---|
| Case 1 | Case 2 | 5% | 10% | 15% | 20% |
| **Significance level = 5%, Power = 95%** | | | | | |
| 10% | 90% | 468 | 133 | 65 | 39 |
| 20% | 80% | 751 | 200 | 93 | 54 |
| 30% | 70% | 947 | 244 | 110 | 63 |
| 40% | 60% | 1056 | 266 | 118 | 65 |
| 50% | 50% | 1077 | 266 | 115 | 63 |
| 60% | 40% | 1012 | 244 | 103 | 54 |
| 70% | 30% | 860 | 200 | 80 | 39 |
| 80% | 20% | 621 | 133 | 46 | NA |
| 90% | 10% | 291 | NA | NA | NA |
| **Significance level = 5%, Power = 90%** | | | | | |
| 10% | 90% | 362 | 102 | 49 | 30 |
| 20% | 80% | 589 | 156 | 72 | 42 |
| 30% | 70% | 746 | 191 | 87 | 49 |
| 40% | 60% | 834 | 210 | 93 | 52 |
| 50% | 50% | 853 | 211 | 92 | 50 |
| 60% | 40% | 804 | 195 | 83 | 44 |
| 70% | 30% | 686 | 161 | 66 | 33 |
| 80% | 20% | 498 | 109 | 40 | NA |
| 90% | 10% | 239 | NA | NA | NA |
| **Significance level = 5%, Power = 80%** | | | | | |
| 10% | 90% | 253 | 69 | 33 | 20 |
| 20% | 80% | 419 | 109 | 50 | 29 |
| 30% | 70% | 534 | 136 | 62 | 35 |
| 40% | 60% | 600 | 151 | 67 | 38 |
| 50% | 50% | 617 | 153 | 67 | 37 |
| 60% | 40% | 583 | 142 | 61 | 33 |
| 70% | 30% | 501 | 119 | 50 | 26 |
| 80% | 20% | 368 | 83 | 32 | NA |
| 90% | 10% | 184 | NA | NA | NA |

Case 1: $H_0$: $P \# P_0$ vs $H_A$: $P > P_0$; Case 2: $H_0$: $P \$ P_0$ vs $H_A$: $P < P_0$; $P = P_1$ at the boundary of the gray region determined in Step 6 of the DQO Process.

**Table 5-3. Sample Size Needed for a One-Sample Test for a Population Proportion, P, at a 10% Significance Level**

| $P_0$ | | $\lvert P_1 - P_0 \rvert$ | | | |
|---|---|---|---|---|---|
| Case 1 | Case 2 | 5% | 10% | 15% | 20% |
| **Significance level = 10%, Power = 95%** | | | | | |
| 10% | 90% | 378 | 109 | 54 | 33 |
| 20% | 80% | 601 | 161 | 75 | 44 |
| 30% | 70% | 753 | 195 | 88 | 50 |
| 40% | 60% | 837 | 211 | 93 | 52 |
| 50% | 50% | 852 | 210 | 91 | 49 |
| 60% | 40% | 798 | 191 | 80 | 42 |
| 70% | 30% | 676 | 156 | 62 | 30 |
| 80% | 20% | 484 | 102 | 34 | NA |
| 90% | 10% | 221 | NA | NA | NA |
| **Significance level = 10%, Power = 90%** | | | | | |
| 10% | 90% | 284 | 81 | 40 | 24 |
| 20% | 80% | 456 | 121 | 57 | 33 |
| 30% | 70% | 575 | 148 | 67 | 38 |
| 40% | 60% | 641 | 161 | 72 | 40 |
| 50% | 50% | 654 | 161 | 70 | 38 |
| 60% | 40% | 615 | 148 | 63 | 33 |
| 70% | 30% | 522 | 121 | 49 | 24 |
| 80% | 20% | 377 | 81 | 28 | NA |
| 90% | 10% | 177 | NA | NA | NA |
| **Significance level = 10%, Power = 80%** | | | | | |
| 10% | 90% | 188 | 53 | 25 | 15 |
| 20% | 80% | 308 | 81 | 38 | 22 |
| 30% | 70% | 392 | 100 | 45 | 26 |
| 40% | 60% | 439 | 110 | 49 | 28 |
| 50% | 50% | 449 | 111 | 49 | 27 |
| 60% | 40% | 424 | 103 | 44 | 24 |
| 70% | 30% | 363 | 86 | 36 | 18 |
| 80% | 20% | 265 | 59 | 22 | NA |
| 90% | 10% | 130 | NA | NA | NA |

Case 1: $H_0$: $P \# P_0$ vs $H_A$: $P > P_0$, Case 2: $H_0$: $P \$ P_0$ vs $H_A$: $P < P_0$; $P = P_1$ at the boundary of the gray region determined in Step 6 of the DQO Process; NA = not

**Table 5-4. Sample Size Needed for a Two-Sample t-Test**

| Significance | | Effect Size | | | | |
|---|---|---|---|---|---|---|
| **Level** | **Power** | **10%** | **20%** | **30%** | **40%** | **50%** |
| 5% | 95% | 2,166 | 542 | 242 | 136 | 88 |
| | 90% | 1,714 | 429 | 191 | 108 | 70 |
| | 80% | 1,238 | 310 | 139 | 78 | 51 |
| 10% | 95% | 1,714 | 429 | 191 | 108 | 69 |
| | 90% | 1,315 | 329 | 147 | 83 | 53 |
| | 80% | 902 | 226 | 101 | 57 | 37 |

Case 1: $H_0$: $:_1 - :_2 \# *_0$ vs $H_A$: $:_1 - :_2 > *_0$; Case 2: $H_0$: $:_1 - :_2 \$ *_0$ vs $H_A$: $:_1 - :_2 < *_0$. In either case, $*_1 = (:_1 - :_2)$ at the boundary of the gray region determined in Step 6 of the DQO Process, and the effect size is $100 * |*_1 - *_0| / \sigma$.

See Table 24.1 of Cohen (1988) for a more extensive tabulation.

**Table 5-5. Sample Size Needed for a Two-Sample Test for Proportions at a 5% Significance Level**

| $P_1$ | | $\lvert P_1 - P_2 \rvert$ | | | |
|---|---|---|---|---|---|
| Case 1 | Case 2 | 5% | 10% | 15% | 20% |
| **Significance level = 5%, Power = 95%** | | | | | |
| 10% | 90% | 947 | 276 | 139 | 87 |
| 20% | 80% | 1510 | 406 | 192 | 114 |
| 30% | 70% | 1900 | 493 | 226 | 130 |
| 40% | 60% | 2116 | 536 | 240 | 136 |
| 50% | 50% | 2160 | 536 | 236 | 130 |
| 60% | 40% | 2030 | 493 | 212 | 114 |
| 70% | 30% | 1727 | 406 | 168 | 87 |
| 80% | 20% | 1250 | 276 | 106 | NA |
| 90% | 10% | 601 | NA | NA | NA |
| **Significance level = 5%, Power = 90%** | | | | | |
| 10% | 90% | 750 | 219 | 110 | 69 |
| 20% | 80% | 1195 | 322 | 152 | 90 |
| 30% | 70% | 1503 | 390 | 179 | 103 |
| 40% | 60% | 1675 | 424 | 190 | 108 |
| 50% | 50% | 1709 | 424 | 187 | 103 |
| 60% | 40% | 1606 | 390 | 167 | 90 |
| 70% | 30% | 1366 | 322 | 133 | 69 |
| 80% | 20% | 990 | 219 | 84 | NA |
| 90% | 10% | 476 | NA | NA | NA |
| **Significance level = 5%, Power = 80%** | | | | | |
| 10% | 90% | 541 | 158 | 80 | 50 |
| 20% | 80% | 863 | 232 | 110 | 65 |
| 30% | 70% | 1086 | 282 | 129 | 75 |
| 40% | 60% | 1209 | 307 | 138 | 78 |
| 50% | 50% | 1234 | 307 | 135 | 75 |
| 60% | 40% | 1160 | 282 | 121 | 65 |
| 70% | 30% | 987 | 232 | 96 | 50 |
| 80% | 20% | 715 | 158 | 61 | NA |
| 90% | 10% | 344 | NA | NA | NA |

Case 1: $H_0$: $P_1 - P_2 \# 0$ vs $H_A$: $P_1 - P_2 > 0$; Case 2: $H_0$: $P_1 - P_2 \$ 0$ vs $H_A$: $P_1 - P_2 < 0$; NA = Not applicable.

**Table 5-6. Sample Size Needed for a Two-Sample Test for Proportions at a 10% Significance Level**

| $P_1$ | | $\lvert P_1 - P_2 \rvert$ | | | |
|---|---|---|---|---|---|
| Case 1 | Case 2 | 5% | 10% | 15% | 20% |
| **Significance level = 10%, Power = 95%** | | | | | |
| 10% | 90% | 750 | 219 | 110 | 69 |
| 20% | 80% | 1195 | 322 | 152 | 90 |
| 30% | 70% | 1503 | 390 | 179 | 103 |
| 40% | 60% | 1675 | 424 | 190 | 108 |
| 50% | 50% | 1709 | 424 | 187 | 103 |
| 60% | 40% | 1606 | 390 | 167 | 90 |
| 70% | 30% | 1366 | 322 | 133 | 69 |
| 80% | 20% | 990 | 219 | 84 | NA |
| 90% | 10% | 476 | NA | NA | NA |
| **Significance level = 10%, Power = 90%** | | | | | |
| 10% | 90% | 575 | 168 | 85 | 53 |
| 20% | 80% | 917 | 247 | 117 | 69 |
| 30% | 70% | 1153 | 299 | 137 | 79 |
| 40% | 60% | 1285 | 326 | 146 | 83 |
| 50% | 50% | 1311 | 326 | 143 | 79 |
| 60% | 40% | 1232 | 299 | 129 | 69 |
| 70% | 30% | 1048 | 247 | 102 | 53 |
| 80% | 20% | 759 | 168 | 64 | NA |
| 90% | 10% | 365 | NA | NA | NA |
| **Significance level = 10%, Power = 80%** | | | | | |
| 10% | 90% | 395 | 115 | 58 | 37 |
| 20% | 80% | 629 | 170 | 80 | 48 |
| 30% | 70% | 792 | 206 | 94 | 55 |
| 40% | 60% | 882 | 224 | 100 | 57 |
| 50% | 50% | 900 | 224 | 98 | 55 |
| 60% | 40% | 846 | 206 | 88 | 48 |
| 70% | 30% | 720 | 170 | 70 | 37 |
| 80% | 20% | 521 | 115 | 44 | NA |
| 90% | 10% | 251 | NA | NA | NA |

Case 1: $H_0$: $P_1 - P_2 \# 0$ vs $H_A$: $P_1 - P_2 > 0$; Case 2: $H_0$: $P_1 - P_2 \$ 0$ vs $H_A$: $P_1 - P_2 < 0$; NA = Not applicable.

# CHAPTER 6

# STRATIFIED SAMPLING

## 6.1 OVERVIEW

Stratified sampling is a sampling design in which prior information about the population is used to determine groups (called strata) that are sampled independently. Each possible sampling unit or population member belongs to exactly one stratum. There can be no sampling units that do not belong to any of the strata and no sampling units that belong to more than one stratum. When the strata are constructed to be relatively homogeneous with respect to the variable being estimated, a stratified sampling design can produce estimates of overall population parameters (for example, mean, proportion) with greater precision than estimates obtained from simple random sampling. Using proportional allocation to determine the number of samples to be selected from each stratum will produce estimates of population parameters with precision at least as good as, and possibly better than, estimates obtained using simple random sampling (regardless of how the strata are defined). However, if optimal allocation is used to assign samples to the strata, and the estimates of the variance within the strata are not close to the actual values, the level of precision in the resulting estimates may be worse than the level of precision for simple random sampling.

Stratified random sampling also is often used to produce estimates with prespecified precision for important subpopulations. For example, one of the most common uses of stratification is to account for spatial variability by defining geographic strata, especially when results need to be reported separately for particular geographic areas or regions. Strata may also be defined temporally. Temporal strata permit different samples to be selected for specified time periods and, hence, also permit designing the sample to support separate estimates for different time periods (for example, seasons) with prespecified precision. Hence, temporally stratified sampling designs support accurate monitoring of trends.

## 6.2 APPLICATION

The method of defining the strata depends on the purpose of the stratification. One of the principal reasons for using a stratified design is to ensure a more representative sample by distributing the sample throughout the spatial and/or temporal dimensions of the population. For instance, a sample drawn with a simple random sample may not be uniformly distributed in space and/or time because of the randomness. Such a sample may not be as representative of the population as a sample obtained by stratifying the study area and independently selecting a sample from each stratum.

Stratification may produce gains in precision in the estimates of population characteristics. If the investigator has prior knowledge of the spatial distribution of the study area, the strata should be

defined so that the area within each stratum is as homogeneous as possible. In addition, the strata can be defined using reliable data on another variable that is highly correlated with the variable to be estimated. If the sample is allocated either proportionally or optimally to the strata, the resulting estimates will have greater precision than if no stratification were used. The variable providing the information used to establish the strata is referred to throughout this chapter as an "auxiliary variable."

Stratification is advisable if a population is subdivided into groups and certain information is desired separately for each group. If estimates (for example, means, proportions, etc.) are desired for particular groups or regions, each group or region would be assigned as a separate stratum. Stratification also is useful if different parts of a population present different sampling issues that may need to be addressed separately. Field conditions may need different sampling procedures for different groups of the population in order to be efficient. This approach is facilitated by stratified sampling because, by definition, each stratum is sampled independently of the other strata. If unbiased estimators of the stratum mean and variance exist for each stratum, then one also can produce unbiased estimates of the overall mean and variance. Field conditions may need different sampling procedures for different groups of the population in order to be efficient. This approach is facilitated by stratified sampling because each stratum can use a different statistical sampling method.

## 6.3 BENEFITS

Stratification can be useful when the implementation of different sampling designs in each stratum could reduce costs associated with the sample selection. The strata can be defined in order to minimize costs associated with sampling at various sites. Study sites that are close in proximity to one another can be assigned to one stratum to minimize the travel time for a team of field personnel to take samples at these locations. Also, if the costs of collecting samples at a portion of a study site are much greater than the rest of the study site, the most costly portion of the site can be assigned as a stratum to minimize sample collection costs. Groups of the population with certain characteristics, which may or may not be the same as the primary stratification variables, can be used as strata in order to ensure that a sufficient number of sampling units appear in the sample for estimates or other analysis of the groups. For example, the investigator may want to stratify the country by average yearly rainfall in order to increase the precision of estimates and may also want to stratify by EPA region to obtain estimates for each region. Stratification can also ensure that certain rare groups of the population that are of interest for estimates or analysis, and that may not otherwise have sufficient sample sizes, have the sample sizes necessary to perform the desired analyses.

When stratification is based on correlation with an auxiliary variable which is adequately correlated with the variable of interest, stratification can produce estimates with increased precision compared with simple random sampling or, equivalently, achieve the same precision with fewer observations. For increased precision, the auxiliary variable used to define the strata should be highly correlated with the outcomes being measured. The amount of increase in precision over simple random

sampling depends on the strength of the correlation between the auxiliary variable and the outcome variable being measured. Consider a situation in which a prior study had found that the amount of clay in the soil is correlated with the amount of a chemical that remains in the soil. In this case, the investigator could use a map of the study area showing the amount of clay in the soil to define the strata needed to estimate the concentration of the chemical. Strata can be defined in order to minimize costs to attain a given level of precision or to maximize precision for a given cost. Example 6-1 shows how the appropriate use of stratification in a planned sampling design can produce estimates with increased precision or need fewer samples as compared to simple random sampling.

## 6.4 LIMITATIONS

Stratified sampling needs reliable prior knowledge of the population in order to effectively define the strata and allocate the sample sizes. The gains in the precision, or the reductions in cost, depend on the quality of the information used to set up the stratified sampling design. Any possible increases in precision are particularly dependent on strength of the correlation of the auxiliary, stratification variable with the variable being observed in the study. Precision may be reduced if Neyman or optimal allocation is used and if the auxiliary variable used for the optimization calculations does not accurately reflect the variability of observations for the study.

As with simple random sampling, with a stratified sampling plan the investigator may encounter difficulties identifying and gaining access to the sampled locations in the field. Such limitations may reduce the expected gains in precision anticipated by using a stratified sampling scheme.

## 6.5 IMPLEMENTATION

### 6.5.1 How do you decide what sample size to use with this design?

The strata should be determined before allocating the sample sizes, and the methods used to define the strata depend on the reasons that stratification is desired. When the strata are to be defined according to an auxiliary variable that is correlated with the variable to be estimated, the optimal definition of the strata is to allocate the strata so that the population included in each stratum is as homogeneous as possible with respect to the auxiliary variable.

Section 5A.6 of Cochran (1977) offers some guidelines on how to optimally assign strata when the auxiliary variable is continuous (i.e., consists of measured values). If the investigator is interested in estimating the overall mean for the population, Cochran suggests defining no more than six strata and using a procedure attributed to Dalenius and Hodges (1959) to determine the optimal cutoff values for each of the strata based on the distribution of the second variable for the population. The steps for determining the Dalenius-Hodges strata are given in Appendix 6-B. Section 5A.7 of Cochran (1977) also provides a discussion and an example of the Dalenius-Hodges procedure. The effectiveness of

using a pilot study to determine the strength of the correlation between the two variables cannot be under estimated.

Once the strata have been defined, a number of options can be used to allocate the sample sizes to each stratum. Equal allocation can be used to assign the same number of samples to be selected within each stratum. Proportional allocation can be used to allocate the samples to the strata so that the proportion of the total sampling units allocated to a stratum is the same as the proportion of sampling units in the population that are classified in that stratum. As mentioned in Section 6.1, proportional allocation can ensure that the precision of the population estimates will be as least as good as, if not better than, the precision without the use of stratification. Optimal allocation has two options:

- Optimize the precision for a fixed study cost.
- Optimize the cost of the study for a fixed level of precision.

If the investigator has a fixed budget in order to collect the samples, the samples could be allocated so that the results would produce the highest precision for the variable to be estimated. If the investigator needs a specific level of precision, the samples could be allocated so that the costs in obtaining the designated level of precision are as low as possible. A special case of the optimal allocation in which the cost of sampling each unit is the same across all strata is Neyman allocation. As previously stated, the extent of the benefits of the stratified sampling design, especially when the optimal sample allocations are used, depend on the quality of the data used to set up the sampling design and the strength of the correlation between the auxiliary variable and the variable to be estimated. However, because the optimal and Neyman sample allocations depend on auxiliary data, the increase (or possible decrease) in precision of the estimates as compared to simple random sampling depends on the accuracy of the variance values used in the sample allocation calculations. Disproportionate allocation may not work well if good estimates of variances are not available. The formulae for the sample size allocations can be found in Appendix 6-A.

### 6.5.2 How do you decide where to take samples with this design?

Once the strata are established, any sampling design can be used to select the samples within each stratum. Where to select these samples will depend on the choice of sampling design that is used (Section 6.6).

## 6.6 RELATIONSHIP TO OTHER SAMPLING DESIGNS

As mentioned earlier, any sampling design can be used within each stratum. The choices include, but are not limited to, simple random sampling, quasi-random sampling, grid sampling, and even another level of stratified sampling.

## 6.7 EXAMPLE

An investigator wants to estimate the average concentration of arsenic in the surface soil around the smoke stack at a hazardous waste incinerator facility to determine if the soil has been contaminated above the naturally occurring concentrations of arsenic for the region. Samples are to be taken within 500 meters from the smoke stack. Information gathered from prior studies indicates that the concentration of arsenic will be higher in the area along the prevailing wind direction and that the variability of the concentration of arsenic in the soil will be higher for clayey soils compared to sandy soils. Because the hazardous waste incinerator facility is located along the ocean coast, the prevailing winds flow from the east. The precision for the estimate of the concentration of arsenic can be increased by dividing the study area into strata according to the prevailing wind direction and the type of soil (see Figure 6-1).

Budget restrictions will only allow 60 samples to be taken from the area around the smoke stack. The study area was stratified according to Figure 6-1, and the Neyman allocation (described in Section 6.5.1) was used to determine the number of samples to be randomly selected within each stratum. The summary statistics for the stratified samples are shown in Table 6-1. Suppose that a simple random sample of 60 soil samples was also taken from the study area for comparison of the performance of the designs. Table 6-1 shows that taking 60 samples by simple random sampling and stratified random sampling produce similar estimates for the mean concentration of arsenic, but the standard error associated with the stratified random sample is lower (i.e., the precision is higher) than that of the simple random sample. Table 6-2 shows that the investigator would have only needed to take 40 soil samples using stratified random sampling in order to get a precision similar to that obtained by analysis of 60 samples taken by simple random sampling. This result is shown by comparing the standard errors and the 95% confidence intervals shown for the various sample sizes under stratified random sampling and simple random sampling. If a particular precision was desired for this study (for example, a standard error of 1.00 for estimating the mean), the investigator could reduce the costs of obtaining an estimate of the average concentration of arsenic by using a stratified sampling design as described above instead of a simple random sampling design.

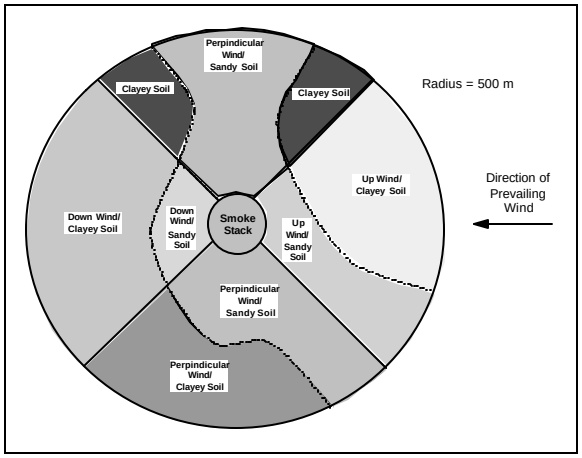


**Figure 6-1. Stratification of Area to Be Sampled**

**Table 6-1. Summary Statistics for Simple and Stratified Random Samples**

| | Simple Random Sampling | Stratified Random Sampling | | | | |
|---|---|---|---|---|---|---|
| | | Down-wind/ Clayey Soil | Down-wind/ Sandy Soil | Perpendicular Wind/Clayey Soil | Perpendicular Wind/Sandy Soil | Overall |
| **# samples** | 60 | 43 | 5 | 10 | 2 | 60 |
| **mean** | 19.81 | 46.16 | 12.66 | 9.49 | 10.20 | 22.94 |
| **standard error** | 4.35 | 9.99 | 4.63 | 2.28 | 3.12 | 3.68 |

**Table 6-2. Number of Samples Needed to Produce Various Levels of Precision for the Mean**

| | Simple Random Sampling | Stratified Random Sampling | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **# samples** | 60 | 60 | 40 | 20 | 14 | 9 | 8 | 7 |
| **standard error** | 4.35 | 3.68 | 4.51 | 6.41 | 7.57 | 9.06 | 9.73 | 10.59 |
| **95% Confid. Interval** | ±8.69 | ±7.36 | ±9.12 | ±10.57 | ±16.35 | ±20.50 | ±22.43 | ±25.04 |

## APPENDIX 6-A

## FORMULAE FOR ESTIMATING SAMPLE SIZE SPECIFICATIONS FOR STRATIFIED SAMPLING DESIGNS

This appendix contains formulae for several commonly used estimates of sample size *n*.

L = number of strata
$N_h$ = total number of units in stratum h
N = total number of units in population, $N = \sum_{h=1}^{L} N_h$
$n_h$ = number of units sampled in stratum h

C To calculate the overall mean and the variance of the overall mean for stratified random sampling:

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$$

$$\text{variance of } \bar{x}_{st} = \sum_{h=1}^{L} \left( w_h^2 (1 - \frac{n_h}{N_h}) \frac{s_h^2}{n_h} \right)$$

where $\bar{x}_h$ is the ordinary mean of stratum h, and $s_h^2$ is the ordinary estimated variance of stratum h.

C To calculate the sample size within the stratum:

n = total number of units sampled, $n = \sum_{h=1}^{L} n_h$
$\sigma_h$ = prior known standard deviation in stratum h
$W_h$ = stratum weight, $W_h = N_h / N$
C = total budget
$C_0$ = initial fixed costs
$C_h$ = cost per sample for stratum h
V = fixed variance

– equal allocation: $n_h = \frac{n}{L}$

– proportional allocation: $n_h = nW_h$

– Neyman allocation: $n_h = n\left(\frac{W_h\sigma_h}{\sum_{h=1}^{L} W_h\sigma_h}\right)$ Note that in practice, $\sigma_h$ is replaced by $s_h$.

– optimal allocation for fixed cost: $n_h = \frac{(C - C_0)W_h\sigma_h/\sqrt{C_h}}{\sum_{k=1}^{L} W_h\sigma_h\sqrt{C_h}}$ Again, in practice, $\sigma_h$ is replaced by $s_h$.

– optimal allocation for a fixed margin of error for each stratum:

$$n = \frac{z^2_{1-\frac{\alpha}{2}}\left(\sum_{h=1}^{L} W_h s_h^2 / d\right)}{1 + z^2_{1-\frac{\alpha}{2}}\left(\sum_{h-1}^{L} W_h s_h^2 / (d^2 N)\right)}$$

where d is the "margin of error" for each estimate within the strata

## APPENDIX 6-B

## DALENIUS-HODGES STRATIFICATION PROCEDURE

This procedure is used to determine the optimal cut-off points for stratification using a variable ($y$) that is highly correlated with the variable of interest. Often this is a continuous variable expected to be highly correlated with the primary outcome to be measured in the study.

1. Form an initial set of $K$ intervals that cover the entire range of observed $y$ values. Let $[A_{i-1}, A_i]$ denote the endpoints of the $i^{th}$ interval ($i=1,2,3,...,K-1$). Count the number of observations, $N_i$, in each interval.

2. Calculate $D_i = A_i - A_{i-1}$ and $T = \sqrt{N_i D_i}$.

3. For each interval $i$, calculate $C_i = \sum_{j=1}^{i} T_j$. That is, add all the $T_j$ from the first interval up to, and including, interval $i$. This makes a cumulative count.

4. Calculate Q = Total/L where Total = $\sum_{i=1}^{L} T_i$ and $L$ is the desired number of strata.

5. For each interval $i$, calculate $C_i/Q$ and round it up to the next higher integer. This now gives the stratum number to which the observations in interval $i$ will be classified.

For example, supply the correlated variable $y$ ranges from 0 to 50, and suppose $L=3$ strata will be created. The Dalenius-Hodges procedure can be used to define the strata:

| Interval | $D_i$ | $N_i$ | $T_i$ | $C_i$ | $C_i/Q$ (Q = 225.3/3=75.1) | Rounded value |
|---|---|---|---|---|---|---|
| 0-5 | 5 | 254 | 35.6 | 35.6 | 0.47 | 1 |
| 5-14 | 9 | 195 | 41.9 | 77.5 | 1.03 | 2 |
| 14-20 | 6 | 160 | 31.0 | 108.5 | 1.44 | 2 |
| 20-30 | 10 | 135 | 36.7 | 145.2 | 1.93 | 2 |
| 30-35 | 5 | 90 | 21.2 | 166.4 | 2.22 | 3 |
| 35-45 | 10 | 155 | 39.4 | 205.8 | 2.74 | 3 |
| 45-50 | 5 | 76 | 19.5 | 225.3 | 3.00 | 3 |
| Total | | 1065 | 225.3 | | | |

It follows that the 1st stratum contains $y$-values 0-5, the second stratum contains $y$-values between 5 and 30, the last stratum contains y-values between 30 and 50.

## APPENDIX 6-C

## CALCULATING THE MEAN AND STANDARD ERROR

Since it would be very difficult to estimate the number of soil samples, $N_h$, which could be taken in each stratum, assign a weight, $W_h$, to each stratum based on the percentage of the study area covered by the stratum. For instance, if down-wind clayey soil covers 35% of the study area, then $W_h$=0.35 for this stratum. Note that the sum of the weights for all strata should equal 1.

Step1: Calculate the sample size, $n_h$, for each stratum with a total sample size of 60 (n=60) under Neyman allocation using the equation:

$$n_h = n \frac{W_h s_h}{\sum_{h=1}^{L} W_h s_h}$$

The assumed population standard deviations, $F_h$, and weights, $W_h$, for each stratum were assigned as follows:

| Stratum | Weight ($W_h$) | Population Standard Deviation, $F_h$ | Neyman Allocation Sample Size, $n_h$ |
|---|---|---|---|
| Down-Wind / Clayey Soil | 0.35 | 75 | 43 |
| Down-Wind / Sandy Soil | 0.15 | 20 | 5 |
| Perpendicular Wind / Clayey Soil | 0.30 | 20 | 10 |
| Perpendicular Wind / Sandy Soil | 0.20 | 5 | 2 |

Step 2: Calculate the mean, $\bar{x}_h$, and variance, $s_h^2$, of the samples within each stratum using the standard formulae used for Simple Random Sampling. The results are summarized in the following table:

| Stratum | Mean $\bar{x}_h$ | Variance $s_h^2$ | Sample Size $n_h$ | Weight $W_h$ |
|---|---|---|---|---|
| Down-Wind/Clayey Soil | 46.16 | 4287.84 | 43 | 0.35 |
| Down-Wind/Sandy Soil | 12.66 | 107.08 | 5 | 0.15 |
| Perpendicular Wind/Clayey Soil | 9.49 | 51.88 | 10 | 0.30 |
| Perpendicular Wind/Sandy Soil | 10.20 | 19.52 | 2 | 0.20 |

Step 3: Calculate the mean, $\bar{x}_{st}$, under stratified sampling

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h = 22.94$$

When N is very large, as it is in this example, the equation for the variance under stratified sampling reduces to:

$$\text{variance of } \bar{x}_{st} = \left[ \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} \right] = 13.55$$

Step 4: The standard error of the stratified sampling mean is the square root of the variance:

$$\text{standard error of } \bar{x}_{st} = \left[ \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} \right]^{1/2} = 3.68$$

# CHAPTER 7

# SYSTEMATIC/GRID SAMPLING

## 7.1 OVERVIEW

Systematic sampling, also called grid sampling or regular sampling, consists of collecting samples at locations or over time in a specified pattern. For example, samples might be collected from a square grid over a set geographical area or at equal intervals over time. Systematic designs are good for uniform coverage, ease of use, and the intuitive notion that important features of the population being sampled will not be missed. Also, samples taken at regular intervals, such as at every node of an area defined by a grid, are useful when the goal is to estimate spatial or temporal correlations or to identify a pattern.

Systematic sampling is used to ensure that the target population is fully and uniformly represented in the set of *n* samples collected. To make systematic sampling a probability-based design, the initial sampling location is chosen at random. Then the remaining (*n*-1) sampling locations are chosen so all *n* are spaced according to some pattern.

There are two major applications for systematic sampling:

- **Spatial designs.** Samples may be collected in one, two, or three dimensions if the population characteristic of interest has a spatial component. Sampling along a line or transect is an example of sampling in one dimension. Sampling every node on a grid laid over an area of interest is sampling in two dimensions. If depth or volume is of interest, samples can be taken at regular grid intervals in three dimensions, such as uniformly spacing samples from a pile of dirt both horizontally and vertically. Several options for systematic two-dimensional sampling in space are shown in Figure 7-1 (Gilbert, 1987). In Figure 7-1a, sample location "A" is randomly assigned and all other sampling locations are then known once the grid is laid down. Note how all the sampling points are an equal distance from each other, thus causing problems if the contamination of interest occurs in some fixed pattern. In Figure 7-1b, location "A" is also

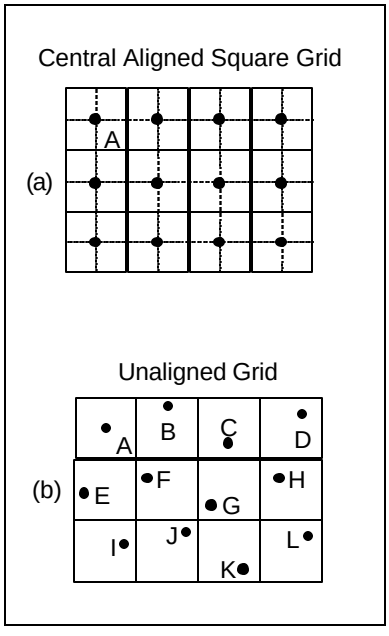


**Figure 7-1. Systematic Designs for Sampling in Space**

selected at random and the remaining locations ("B" through "L") within their square cells are determined randomly within each grid cell. This design has the advantages of randomness combined with good coverage (somewhat similar to the concept of quasi-randomness as discussed at the end of Section 5.5.2).

C **Temporal (periodic) designs.** When samples are selected to represent a target population that changes over time, data collectors would use a one-dimensional sample where every $k^{th}$ unit is selected or a sample is collected at specific points in time. Figure 7-2 (Gilbert, 1987) shows an example of periodic sampling. In this figure, a systematic sample of $n = 4$ units is desired from a finite population of N = 15 units, representing 15 units of time. The 15 units are displayed as a circle for illustration, as if the units were on a clock. The systematic interval between units was determined by computing N/n = 15/4 = 3.75, which is rounded up to 4. Then a random number between 1 and 15 was selected; namely 7. Hence, sampling starts at the $7^{th}$ unit and every $4^{th}$ unit from that point is selected.

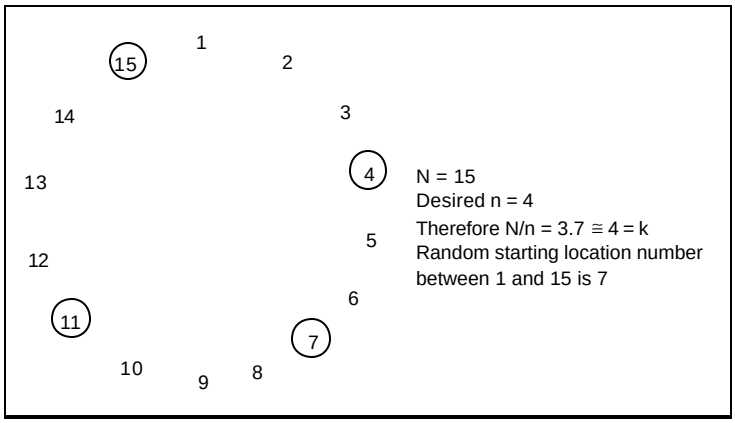


**Figure 7-2. Choosing a Systematic Sample of $n = 4$ Units from a Finite Population of N = 15 Units**

Grid designs can vary in their shape, orientation, and selection criteria for the initial grid node. This flexibility, the intuitive appeal, and easily explained protocol for taking regular samples make systematic sampling one of the more popular and defensible sampling designs.

## 7.2 APPLICATION

Systematic sampling is often used in environmental applications because it is practical and convenient to implement in the field. It often provides better precision (i.e., smaller confidence intervals, smaller standard errors of population estimates) and more complete coverage of the target population than random sampling. Systematic sampling is appropriate if either of the following conditions pertain:

- There is no information about a population and the objective is to determine if there is a pattern or correlation among units, or

- There is a suspected or known pattern or correlation among units at the site and the objective is to estimate the shape of the pattern or the strength of the correlation.

Systematic sampling designs are used in three situations:

1. When making an inference about a population parameter such as the mean when environmental measurements that are known to be heterogeneous. A systematic design is only one of many sampling designs that may be used for making an inference about a population parameter. However, if the concentrations over space or time in the target population are correlated so that the data show definite spatial or temporal patterns, then systematic sampling will often be more efficient (provide a more precise answer for a given amount of sampling) than random sampling. Many automatic samplers use systematic sampling due to the mechanical necessity of taking samples at fixed intervals.

2. When estimating a trend or identifying a spatial or temporal correlation. A systematic design is well suited for this type of problem because a constant distance or time interval between sampling locations or times allows for the efficient estimation of trends and patterns over time or space, as well as the correlation structure needed for modeling. Random sampling would typically need more samples to achieve the same amount of information about the patterns and correlation.

3. When looking for a "hot spot" or making a statement about the maximum size object that could be missed with a given sampling design. If a systematic square, rectangular or triangular grid design is laid over a study site, then it is possible to determine the probability that any size of an approximately elliptical region of elevated concentration ("hot spot") will be hit by a sampling point on the grid. One can also determine the spacing between sampling locations needed to hit an elliptical target with specified probability.

If distinct features exist at a site, such as an ecological cluster or a groundwater plume, then collecting data on a regular grid is the most efficient approach to ensuring such features are actually detected. However, if the scale of the pattern or feature of interest is smaller than the spacing between sampling locations, then the systematic pattern of sampling is not an efficient design unless the spacing between sampling locations is reduced or some other procedure such as composite sampling is introduced into the design.

Systematic sampling would be inappropriate if a known pattern of contamination coincides with the regularity of the grid design. Such a coincidence would result in an overestimation or underestimation of a particular trait in the target population of interest. For example, suppose a line of trees resulted in soil mounds with high contamination along the tree line and a grid line was aligned with the tree line. Then, a decision about the average contamination over an entire area would be upwardly biased by so many samples collected in the high concentration area along the tree line. If prior information is available on the possible patterns of contamination, this information may be important in selecting grid spacing, grid orientation, and whether or not systematic sampling designs have an advantage over other designs.

**What are some more advanced findings on systematic sampling**?

Section 8.2 of Cochran (1977) states that systematic sampling can be considerably more precise than simple random sampling or even stratified random in some situations. He states: "Systematic sampling is more precise than simple random sampling if the variance within the systematic samples is larger than the population variance as a whole. Systematic sampling is precise when units within the same sample are heterogeneous and is imprecise when they are homogeneous." Cochran demonstrates that systematic sampling is capable of providing enhanced performance over other designs depending on the properties of the target population. He provides results from a study of 13 different data sets from natural populations showing a consistent gain in precision using systematic sampling.

Section 8.3 of Gilbert (1987) also discusses the relative performance of systematic sampling for the following types of population structure:

- Populations in random order
- Populations with linear trends
- Populations with periodicities
- Populations with correlations between values in close proximity

Two observations can be made. First, for populations in random order, systematic sampling offers convenience. An example of a random order population might be radioactive fallout from atmospheric nuclear weapons tests that is uniformly distributed over large areas of land. Second, if the population consists entirely of a linear trend, systematic sampling will, on the average, give a smaller variance of $\overline{x}$ (sampling error of the sample mean) than simple random sampling. However, stratified random sampling will, on average, give a smaller variance of $\overline{x}$ than either systematic sampling or simple random sampling.

A comprehensive study by Yfantis, Flatman, and Behar (1987) discusses the level of efficiency and accuracy of different grid types. They conclude that an equilateral triangular grid works slightly

better for the majority of the cases they studied. However, this study did not include the effects of a second or additional phases of sampling. It is possible that when a multiple time period or phased sampling design is planned, the specific type of first-phase sampling grid may be less important than using geostatistical techniques (such as geostatistical simulations) to place second-phase samples in locations that most reduce probabilities of estimation errors (EPA, 1996b).

## 7.3 BENEFITS

Systematic/grid sampling has the following benefits:

- Uniform, known, complete spatial/temporal coverage of the target population is possible. A grid design provides the maximum spatial coverage of an area for a given number of samples.

- The design and implementation of grids is relatively straightforward and has intuitive appeal; field procedures can be written simply. Once an initial point is located, the regular spacing allows field teams to easily locate the next sampling point, except for unaligned or random samples within the grid structure.

- Multiple options are available for implementing a grid design. Often, sampling programs are executed in phases. The initial phase uses broad-scale grids to look for any kind of activity or hit. Once the general area or time frame of the activity of interest has been identified, smaller-scale grids are used to refine the estimates. Alternatively, during a single phase, the total area can be subdivided into areas based on the likelihood of finding properties of interest and different grid spacings used in each sub-area. In addition, one can overlay multiple grids, orient multiple grids in opposite directions, intermix fine-mesh grids with large-mesh grids, and still maintain the constant spacing desired for certain applications, such as estimating the correlation function (i.e., variogram). Standard formulae for estimating sample size and population parameters are adjusted to account for these variations.

- Regularly spaced or regularly timed samples allow for spatial and temporal correlations to be calculated, assuming the pattern of interest is larger than the spacing of the sampled nodes. If correlation over space or time may be present and there are distinct features or patterns in the population to be sampled, constant spacing of samples is often a good option for estimating the features and making predictions of unsampled areas.

- Grid designs can be implemented with little to no prior information about a site. The only inputs needed are the total area to be covered and the number of samples (or

alternatively, the grid spacing) to be used. Grid sampling is often used for pilot studies, scoping studies, and exploratory studies using the assumption that there are no patterns or regularities in the distribution of the contaminant of interest.

Many studies have been performed using simulated data sets to compare the efficiency of alternative sampling designs. All such studies conclude that the overall performance of the design is influenced as much by particular features in the population to be sampled along with the estimators used for estimating population parameters of interest as the type of design chosen.

**What are the results from some more advanced studies?**

In a study on trace elements in contaminated soil to assess the impact of contaminated soil on the environment and on agricultural activities, Wang and Qi (1998) found that given a certain sampling density, systematic sampling had better estimation performance than either a stratified or a random sampling design.

In a study on assessing the percentage cover of crop residue to estimate soil erosion, Li and Chaplin (1995) found that systematic sampling was more precise than random sampling for both corn and soybean residue in most cases. Crop residue is plant material left on the field surface after harvest. Measuring the crop residue cover on the soil surface is essential in the management of soils to reduce erosion. Li and Chaplin laid grid frames on top of a picture taken of fields with corn and soybean residue. The image was then read into a computer program that randomly changed the position of the grid on the picture. Light densities recorded the reading of coverage at each node. The grid design compared favorably to a design where random locations were sampled for coverage readings, using the same number of sampling points as used in the systematic sampling.

In another study, Li and Chaplin (1998) considered both one- and two-dimensional sampling designs for estimating crop residue coverage. Although widely used, no rigorous study exists on the precision of the line transect method. Li and Chaplin used a computer-generated virtual field surface and applied various sampling designs. They found the square grid was more precise than the line transect methods because of the smaller coefficient of variation over a wide range of sampling points and residue cover.

## 7.4 LIMITATIONS

Systematic/grid sampling may not be as efficient as other designs if prior information is available about the population. Such prior information could be used as a basis for stratification or identifying areas of higher likelihood of finding population properties of interest.

If the population properties of interest are aligned with the grid, systematic/grid sampling raises the possibility of an overestimation or underestimation (bias) of a population characteristic. Caution should be used if there is a possibility of a cyclical pattern in the unit or process to be sampled that might match the sampling frequency. For example, one would not want to take air samples every Monday morning if a nearby plant always pressure-cleaned the duct work on Monday morning.

As mentioned earlier, a single systematic sample cannot be used to get a completely valid estimate of the standard error of the mean, i.e., variance of the mean, without some assumptions about the population. This could result in an inaccurate calculation for the confidence interval of the mean. Several approximate methods have been proposed by Wolter (1984) and illustrated in Section 8.6 of Gilbert (1987). One option is to take multiple sets of systematic samples, each with a randomly determined starting point, and calculate an empirical estimate of the standard error of the mean. The use of multiple sets of systematic samples has to be balanced against the cost or feasibility of using the sampling designs incorporating compositing. Methods for estimating the variance of the mean developed for simple random sampling plans can be used with confidence only when the population is in random order.

## 7.5 IMPLEMENTATION

Systematic sampling designs are relatively straightforward to implement. You need to know how many samples to take and where to take them.

### 7.5.1 How do you decide how many samples to take?

Many of the sample size formulae provided for simple random sampling (i.e., the sample size formula for estimating a mean provided in Chapter 4) can be used for systematic sampling as long as there are no strong cyclical patterns, periodicities, or significant spatial correlations between pairs of sample locations not intoduced as part of the grid or systematic process. For the hot spot problem, there are nomographs provided in Section 10.1 of Gilbert (1987) and a computer program called ELIPGRID PC (Davidson, 1995) for calculating the optimal grid spacing for a hot spot of prespecified size and shape with a specified confidence of finding the hot spot. Li and Chaplin (1998) discuss how to design grid sampling patterns with the least number of sampling points to achieve a specified precision based on results.

### 7.5.2 How do you decide where to take samples?

There are many variations on patterns for regular spacing of systematic samples. Patterns include square, rectangles, triangles, circles, and hexagons. Basic geometry can be used to determine internodal spacing. For example, for the two-dimensional sampling problem, EPA has detailed guidance on how to locate samples using a systematic sampling design (EPA, 1989). Figure 7-3, taken

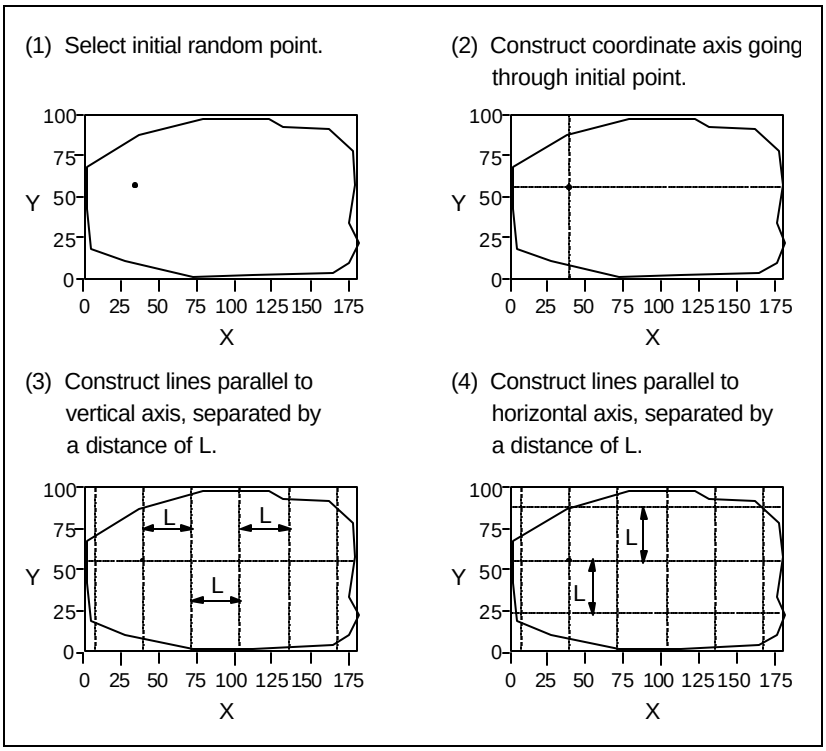


**Figure 7-3. Locating a Square Grid Systematic Sample**

from that document, summarizes how to lay out a square grid. Once a sample size $n$ and the area $A$ to be sampled have been specified, Equations 7-1 and 7-2 can be used to calculate the spacing between adjacent sampling locations. For the square grid, the distance L between the vertical and horizontal parallel lines is:

$$L = \sqrt{\frac{A}{n}} \tag{7-1}$$

For the triangular grid, the distance L becomes:

$$L = \sqrt{\frac{A}{0.866n}} \tag{7-2}$$

For one-dimensional sampling, the procedure theoretically is even simpler, but the complexities for the one-dimensional problem come in the application. For example, the line transect method is used extensively by U.S. Department of Agriculture technicians as a quick means to estimate agricultural conditions, such as plant coverage. To conduct a measurement in a certain area, a cord with 50 to 100 equally spaced beads is stretched diagonally across the crop rows. Using the same point on each bead—for example, the leading edge—those beads are counted that have the plant characteristic of

interest under them when viewed directly from above. This count is divided by the total number of beads on the cord to give an observation of the percent occurrence. An average of three to five observations in the area is used to estimate field totals. The transect length, size of the cord, and marker spacing are part of the protocol.

For more discussion of the diagonal line transect method, refer to the MidWest Plan Service (MWPS, 1992). Also, see Li and Chapin (1998) for more detailed information on implementing this method.

## 7.6 RELATIONSHIP TO OTHER SAMPLING DESIGNS

Systematic sampling can be used in place of random sampling in many of the designs discussed in this document. For example, sampling on a grid pattern can be conducted within each stratum of a stratified sampling plan (Chapter 5). The key criteria for using a systematic design is that a random starting location be identified for the selection of the initial unit and the grid layout cannot coincide with a characteristic of interest in the population.

For example, the Environmental Monitoring and Assessment Program uses a sampling strategy that has multiple stages and involves aspects of stratified and systematic sampling. The first stage of the design is a triangular grid covering the conterminous United States. The grid is randomly situated over the U.S. land mass; the interpoint distance along the grid is approximately 27 kilometers, and the ratio of area to number of grid points is approximately 635 square kilometers per grid point. The grid design is good for measuring those ecological resources that do not change position over the time of the survey and that need to be sampled repeatedly over time. The multistage design permits the design to be tailored to the resources of interests and purposes of the reporting. During the first stage, data may be collected at random sample grid points; on the basis of these data, informed choices can be made for the definition, stratification, and so on of second and lower stage units. In preparation for the second stage, a randomly placed hexagonal template is constructed over the region. The typical size of the template is 16 hexagons per grid point (Cox et al.,1995).

The combination of systematic and random sampling was demonstrated in a study by Cailas et al. (1995) in proposing a methodology for an accurate estimation of the total amount of materials recycled. One objective of this comprehensive study of the recycling infrastructure in Illinois was to make an accurate estimation of the amount of total material recycled. It was found that responses from a small number of previously identified critical facilities were essential for an accurate estimation of the total amount of material recycled. The combined design consisted of systematically sampling the critical facilities and randomly sampling the remaining ones. This application yielded an accurate estimate with less than 1% difference from the actual amount recycled. This was done with only 15% of the total number of recycling facilities included in the critical facilities subpopulation.

## 7.7 EXAMPLES

### 7.7.1 Implementing Triangular Sampling

This example is taken from EPA (1989, 1992). Suppose 30 samples were to be taken from an area of 14,025 $m^2$. This area is shown in gray in Figure 7-4.

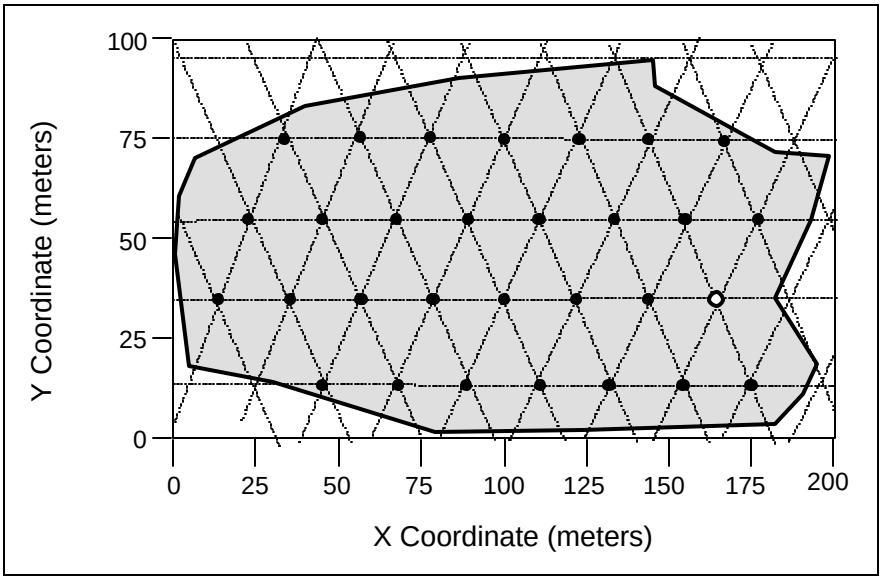


**Figure 7-4. Map of an Area to Be Sampled Using a Triangular Sampling Grid**

The following steps are performed:

1. The boundaries for the problem are determined to be $X_{min} = 0$, $Y_{min} = 0$, $X_{max} = 200$, and $Y_{max} = 100$.
2. A random number generator is used to obtain two random numbers (call them $R_1$ and $R_2$) between 0 and 1. For purposes of this example, the two numbers drawn were 0.820 and 0.360.
3. The random start location (X, Y) is obtained by using the formulae

$$X = X_{min} + R_1(X_{max} - X_{min})$$
$$Y = Y_{min} + R_2 (Y_{max} - Y_{min})$$

4. Substituting the values from Steps 1 and 2 results in the location (X=164, Y=36). This point is shown as an open circle in Figure 7-4.
5. Use the formula for *L* to get:

$$L = \sqrt{\frac{A}{0.866n}}$$

$$L = \sqrt{\frac{14{,}025}{0.866 \times 30}} = 23.23 \ .23$$

6. A line parallel to the x-axis through the point (164,36) is drawn; points are marked off 23 meters apart from this line as shown in Figure 7-4.
7. The midpoint between the last two points along the line is found and a point is marked at a distance (0.866 x 23) = 19.92 (i.e. 20) meters perpendicular to the line at that midpoint. This point is the first sample location on the next line.
8. Points at distance L=23 meters apart are marked on this new line.
9. Steps 6 and 7 are repeated until the triangular grid is determined.

There are now exactly 30 locations marked off in a triangular pattern. In some instances, due to irregular boundaries, it may not be possible to obtain the exact number of samples planned for.

### 7.7.2 Soil Contamination Applications

For applications where the goal of sampling is to evaluate the attainment of cleanup standards for soil and solid media, EPA guidance (EPA, 1992) recommends collecting samples in the reference areas and cleanup units on a random-start equilateral triangular grid except when the remedial-action method may leave contamination in a pattern that could be missed by a triangular grid; in this case, unaligned grid sampling is recommended. There are also many applications for grid sampling when the goal is site characterization. Grid sampling insures all areas are represented in the sample and can provide confidence that a site has been fully characterized.

### 7.7.3 Ecological and Environmental Survey Applications

The National Stream Survey and EPA's Environmental Monitoring and Assessment Program are two large-scale environmental surveys that use variable probability, systematic sampling and a special estimator called the Horvitz-Thompson estimator (Cochran, 1977) to estimate population parameters of ecological interest. For the National Stream Survey, all streams represented as blue lines on 1:250,000 topographic maps define the target population of streams. Sampling units were selected using a square grid, with density of 1 grid node per 64 square miles, imposed on 1:250,000 topographic maps of a target area. A target stream reach was selected into the sample if a grid node fell into the direct watershed of that reach. This protocol resulted in reaches being sampled with probability proportional to direct watershed area. In the Environmental Monitoring and Assessment Program, one objective is to estimate the current condition of the nation's ecological resources on a

regional basis with known confidence. The Environmental Monitoring and Assessment Program's sampling design is based on a systematic, triangular grid (also see discussion in Section 7.6). The grid is used to select a sample in a manner analogous to the National Stream Survey. For example, for sampling lakes, each lake is identified by its "center" and a grid node identifies a lake to be included in the sample as the lake that has a center closest to the grid node. The probability of sampling a given lake is proportional to the area of the polygon enclosing the region closer to that lake's center than to any other lake's center. Larger lakes have a higher probability of being included in the sample (Stehman and Overton, 1994).

When estimating abundance for various animals, samples are often taken along a transect at regular intervals. This is a form of grid sampling. A pronghorn (antelope) abundance study evaluated the efficiency of systematic sampling versus simple random sampling versus probability proportional to size sampling (Kraft et al., 1995). The total number of pronghorn was already known; this was a simulation study to evaluate alternative sampling plans. The sampling unit was a 0.8-km-side linear transect variable in length according to size and shape of the study area. Six different study areas were used. A plane flew along the transect and when a pronghorn was sighted, the pilot circled until the herd could be counted. The goal was to estimate total abundance of pronghorn in an area. For the systematic sampling, the sampling units (transects of different lengths) in an area were numbered; after the first unit was randomly chosen, every $p^{th}$ unit following was selected. For this study, it was found that stratification combined with accurate estimates of optimal stratum sample sizes increased precision, reducing the mean coefficient of variation from 33 without stratification to 25 with stratification. Cost, however, increased with stratification by 23%.

### 7.7.4 Groundwater Applications

For sampling groundwater in fixed wells over time, a systematic sample in time is usually preferred over a simple random sample in time. There are several reasons for this preference: extrapolating from the sample period to future periods is easier with a systematic sample than a simple random sample; seasonal cycles can be easily identified and accounted for in the data analysis; a systematic sample will be easier to administer because of the fixed schedule for sampling times; and most groundwater samples have been traditionally collected using a systematic sample, making comparisons to background more straightforward.

EPA guidance on groundwater sampling for evaluating attainment of cleanup standards (EPA,1992) suggests a variation of systematic sampling when periodic seasonal variations or other repeated patterns are suspected. Several variations are described and recommended depending on the sampling goal as biased estimates may result unless the systematic sample has a spacing small enough to characterize both high and low concentrations. For example, the goals described include identifying or characterizing the pattern of contamination in an aquifer, obtaining comparable period-to-period samples, and making comparisons to background when there are large seasonal fluctuations in the data.

### 7.7.5 Geostatistical Applications

When there is spatial or temporal dependence, moving from one point to another nearby location usually results in values that do not change dramatically. Samples close together will tend to have more similar values than samples far apart. This is often the case in an environmental setting. The method chosen to estimate an overall site mean, as well as the site variance, must properly account for the pattern of spatial continuity. Any non-random or partially random sampling scheme (including a systematic grid design) will tend to produce biased estimates if not adjusted for the degree of spatial correlation. There exist techniques to minimize the biasing impact of spatial correlation while generating reasonable estimates of the mean.

EPA has produced guidance for geostatistical soil sampling (EPA, 1996b). Sampling in support of geostatistical analysis is an important topic and discussed in detail in this EPA document. One important component of geostatistics is the variogram. The variogram is a plot of the variance of paired sample measurements as a function of the distance between samples. Samples taken on a regular grid are desirable for estimating the variogram. While all regular grids tend to work reasonably well in geostatistical applications, there are differences in efficiency depending on the type of grid pattern chosen. The most common grid types include square, triangular, and hexagonal patterns. Entz and Chang (1991) evaluated 16 soil sampling schemes to determine their impact on directional sample variograms and kriging. They concluded that for their case study, grid sampling needs more samples than stratified random sampling and the stratified-grid design, but the accuracy of the kriged estimates was comparable for all sampling designs. They also found that the variograms that were estimated from sample data collected from stratified and grid designs led to the same conclusion about the spatial variability of the soil bulk density (the subject of the study).

### 7.7.6 Hot Spot Problem Application

One application for using grid sampling that is widely encountered in environmental settings is in the spatial context of searching for hot spots. The problem can be formulated several ways:

- What grid spacing is needed to hit a hot spot with specified confidence?
- For a given grid spacing, what is the probability of hitting a hot spot of a specified size?
- What is the probability a hot spot exists when no hot spots were found by sampling on a grid?

For this application, sampling over a gridded area at the nodes is used to search for an object(s) of interest or, alternatively, to be able to state that an object of a specified size cannot exist if a grid node was not intersected. For example, the sampling goal may be to find if at least one 55-gallon drum is buried in an area. Optimal grid spacings for the hot spot problem have been worked out for a range of relative object sizes and orientations. The hot spot problem is discussed extensively in

Chapter 10 of Gilbert (1987). In most situations the triangular grid is more efficient at detecting hot spots than the square or rectangular grid designs.

In summary, if nothing is known about the spatial characteristics of the target population, grid sampling is efficient in finding patterns or locating rare events unless the patterns or events occur on a much finer scale than the grid spacing. If there is a known pattern or spatial or temporal characteristic of interest, grid sampling may have advantages over other sampling designs depending on what is known of the target population and what questions are being addressed by sampling.

# CHAPTER 8

# RANKED SET SAMPLING

## 8.1 OVERVIEW

This chapter describes and illustrates ranked set sampling, an innovative sampling design originally developed by McIntyre (1952). The unique feature of ranked set sampling is that it combines simple random sampling with the field investigator's professional knowledge and judgment to pick places to collect samples. Alternatively, on-site measurements can replace professional judgment when appropriate. The use of ranked set sampling increases the chance that the collected samples will yield representative measurements; that is, measurements that span the range of low, medium, and high values in the population. This results in better estimates of the mean as well as improved performance of many statistical procedures such as testing for compliance with a risk-based or background-based (reference-based) standard. Moreover, ranked set sampling can be more cost-efficient than simple random sampling because fewer samples need to be collected and measured.

The use of professional judgment in the process of selecting sampling locations is a powerful incentive to use ranked set sampling. Professional judgment is typically applied by visually assessing some characteristic or feature of various potential sampling locations in the field, where the characteristic or feature is a good indicator of the relative amount of the variable or contaminant of interest that is present. For example, the relative amounts of a pollutant in randomly selected sampling spots may be assessed based on the degree of surface or subsurface soil staining, discoloration of soil, or the amount of plant defoliation in each spot. Similarly, the yield of a plant species in randomly selected potential 1 meter by 1 meter field plots may be visually assessed based on the density, height, or coloration of vegetation in each plot. This assessment ranks the visually assessed locations from smallest to largest with respect to the variable of interest; it is then used as described in this chapter to determine which spots to actually sample.

In some situations, a more accurate assessment of the relative amounts of a pollutant present at field locations can be provided by an inexpensive on-site measurement. Indeed, the sensitivity and accuracy of in-situ detectors has increased greatly in recent years. Some examples include the following:

- Using ultraviolet fluorescence in the field to measure (screen) for BTEX (benzene, toluene, ethyl benzene, and xylene) and PAHs (polyaromatic hydrocarbons) in soil.
- Using X-ray fluorescence in the field to measure lead or other metals in soil.
- Using total organic halide (TOX) measurements of soil as a screening measurement for volatile organic solvents.

- Using remotely sensed information (aerial photographs and/or spatially referenced databases as found in a geographic information systems) to identify locations to be studied.
- Using distance along a pipeline (longer distance implying lower levels of a contaminant) to approximate the relative concentrations of a contaminant at various distances.

A simple ecological example will illustrate the ranked set sampling approach (based on Stokes and Sager, 1988); a more detailed lead contamination example follows in Section 8.2. The recommended step-by-step process for setting up an ranked set sampling design is presented in Appendix 8-A. Suppose the average individual volume of the trees on a property needs to be estimated. Begin by randomly selecting two trees and judge by eye which tree has the most volume. Mark the *smaller* tree to be carefully measured for volume and ignore the other tree. Next, randomly select another two trees. Mark the *larger* of these two trees and ignore the other tree. Then repeat this procedure, alternatively marking the smaller of the first two trees, then the larger of the second two trees. Repeat this procedure a total of 10 cycles for a total of 40 trees. Twenty of the trees will have been marked and 20 ignored. Of the 20 marked trees, 10 are from a stratum of generally smaller trees and 10 are from a stratum of generally larger trees. Determine the volume of each of the 20 marked trees by careful measurement and use that measurement to estimate the average volume per tree on the lot. In this illustration there were 10 cycles and 2 trees marked per cycle. In practice, the number of trees marked per cycle (the "set size") and the number of cycles is determined using a systematic planning process, as illustrated in Appendix 8-A.

**Example of Using Ranked Set Sampling to Estimate The Mean Lead Concentration in Soil**

Suppose a future residential area is suspected of having lead concentrations in surface soil that exceed background concentrations. As part of the risk assessment process, the soil of the area will be sampled to estimate the mean lead concentration. Prior studies have shown that x-ray fluorescence (XRF) measurements of lead in soil obtained using a hand-held in-situ detector closely correlate with laboratory measurements of lead in soil at the same locations. Furthermore, it was determined that the cost of taking the XRF measurements in the field was very low compared to the cost of laboratory measurements for lead. (Cost considerations are discussed in Appendix 8-A.) Hence, ranked set sampling was selected for data collection instead of simple random sampling (see Appendix 8-A for guidance on how to determine if ranked set sampling is preferred over simple random sampling).

Suppose the systematic planning process employed determined that $n = 12$ soil samples should be collected and measured for lead in the laboratory in order to meet the acceptance and performance criteria for this study (i.e., to have 95% confidence that the estimated mean computed using laboratory lead measurements would be within 25% of the true mean). Also, in order to obtain information to properly compute the variance of this estimated mean, the following replication process was used to obtain the 12 samples. Specifically, $m = 3$ field samples (the "set size") were collected in each of $r = 4$

cycles to obtain the necessary $n = m \times r = 3 \times 4 = 12$ samples that will be measured for lead in the laboratory. A method to determine $m$ and $r$ is provided in Appendix 8-A.

The ranked set sampling method for determining the three field locations to be sampled is as follows:

1. Use simple random sampling to randomly select $m^2 = 3^2 = 9$ locations on the property. Randomly divide the nine locations into $m$ sets of size $m$ (3 sets of size 3). In Figure 8-1 the first set of three locations is denoted by "Set 1," the second set by "Set 2," and the third set by "Set 3."

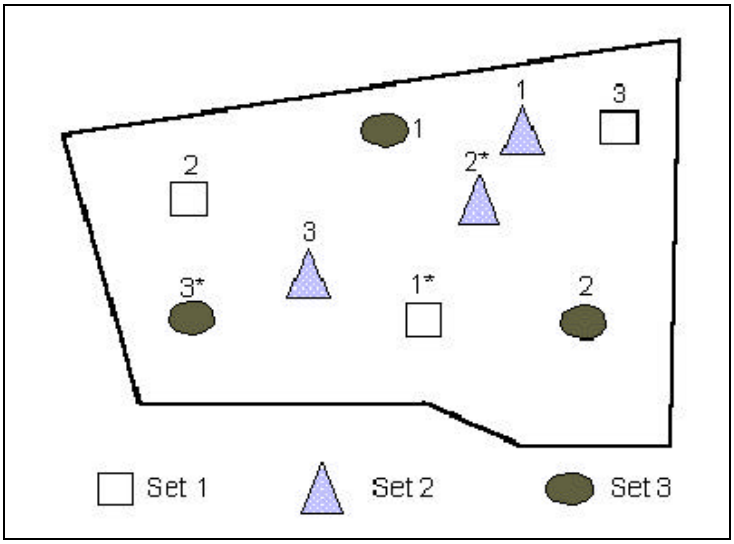


**Figure 8-1. Using Ranked Set Sampling to Select Three Locations**

2. Consider the three locations in Set 1. Make an XRF measurement at each of those three locations and label the locations 1, 2, and 3 to indicate the smallest, middle, and largest XRF measurement, respectively. Collect the first soil sample at location label 1 in Set 1; this location has the smallest XRF lead measurement in Set 1 (labeled 1* in Figure 8-1).

3. Consider the three locations in Set 2 and make an XRF measurement at each of those locations. Collect the second soil sample at label 2 in Set 2; this location has the second highest XRF measurement in Set 2 (labeled 2* in Figure 8-1).

4. Consider the three locations in Set 3 and make an XRF measurement at each of the three locations in that set. Collect the third soil sample at label 3 in Set 3; this location has the highest XRF measurement in Set 3 (labeled 3* in Figure 8-1).

Thus, nine in-situ XRF measurements are used to guide the selection of three soil samples that will be measured for lead in the laboratory. Then, this procedure is repeated $r = 4$ times to obtain the entire $n = m \times r = 3 \times 4 = 12$ soil samples needed. This replication process is needed to estimate the variance of the estimated mean (see Appendix 8-A for the computational formula). In practice, if professional judgment is used to rank the locations in each set, the set size (m = 3 in this example) should be between 2 and 5. Larger values of $m$ make it more difficult to accurately rank the locations

within each set. However, set size larger than five may be practical if field locations are ranked using screening measurements. In general, larger set sizes when using screening measurements are desirable because they result in more precise estimates of the mean.

Note that the above example is a *balanced* ranked set sampling design, that is, the same number of field locations, $r = 4$, are sampled for each of the $m = 3$ ranks. That is, in the above example, a sample is collected at each of four locations expected to have a relatively small value of the variable of interest (lead), as well as at four locations expected to have a mid-value of lead and at four locations expected to have a relatively large value of lead. Unbalanced ranked set sampling designs can also be used, as discussed in Section 8.5.2 and Appendix 8-A.

## 8.2 UNDER WHAT CONDITIONS IS RANKED SET SAMPLING APPROPRIATE?

Ranked set sampling is appropriate when the following conditions hold:

- The cost of laboratory measurements is high relative to the cost of using screening measurements or professional judgment in the field to determine the relative magnitudes of contamination in randomly selected field plots.

- Professional judgment or on-site measurements can accurately determine the relative magnitudes of contamination among randomly selected field locations.

- A more precise estimate of the mean or a more powerful test for compliance is needed than can be achieved for a fixed budget if simple random sampling were used in place of ranked set sampling.

A process whereby costs and accuracy of ranking field locations is considered in setting up a ranked set sampling design is provided in Appendix 8-A.

## 8.3 BENEFITS

A major benefit of ranked set sampling is that it will yield a more precise estimate of the mean than if the same number of measurements is obtained using simple random sampling (McIntyre, 1952; Gilbert, 1995; Johnson et al., 1996; Muttlak, 1996). Table 8-1 illustrates this for the normal distribution with moderate coefficients of variation (CV). For example, suppose the distribution of the variable of interest is normal with a true mean of 1 and a coefficient of variation (CV = standard deviation divided by the mean) of 0.50. Furthermore, suppose our goal is to obtain enough laboratory measurements to have 95% confidence that the estimated mean is within 25% of the true mean. Table

**Table 8-1. Comparing the Number of Samples for Laboratory Analysis Using Ranked Set Sampling***

| Coefficient of Variation (CV)** | Ranked Set Sampling Set Size (m) | Specific Precision of the Estimated Mean with 95% Confidence | | |
|---|---|---|---|---|
| | | 10% | 15% | 25% |
| 0.50 | Simple Random Sampling | 97 | 43 | 16 |
| | Ranked Set Sampling - 2 | 66 | 30 | 12 |
| | Ranked Set Sampling - 3 | 51 | 24 | 9 |
| | Ranked Set Sampling - 5 | 35 | 20 | 10 |
| 0.707 | Simple Random Sampling | 193 | 86 | 31 |
| | Ranked Set Sampling - 2 | 132 | 60 | 22 |
| | Ranked Set Sampling - 3 | 102 | 45 | 18 |
| | Ranked Set Sampling - 5 | 70 | 35 | 15 |
| 1.0 | Simple Random Sampling | 385 | 171 | 62 |
| | Ranked Set Sampling - 2 | 262 | 118 | 42 |
| | Ranked Set Sampling - 3 | 201 | 90 | 33 |
| | Ranked Set Sampling - 5 | 140 | 65 | 25 |

* Adapted from Table 1 in Mode et al. (1999). Table values derived assuming there are no errors in ranking field locations.

** Coefficient of Variation = standard deviation divided by the mean.

8-1 indicates that simple random sampling will need 16 samples, but if ranked set sampling is used with a "set size" of 2, then only 12 samples are needed, reducing sampling and laboratory costs by 25%. If the cost of using professional judgment or on-site measurements is considerably less than the cost of laboratory measurements, then there is a strong motivation to use ranked set sampling rather than simple random sampling. Note in Table 8-1 that when high precision in the estimated mean is needed, the number of samples needed is dramatically reduced as the set size increases. Ranked set sampling also has several other benefits, as follows:

- The estimated mean of ranked set sampling data is a statistically unbiased estimator of the true mean (as is that of a simple random sample).

- Ranked set sampling provides increased ability to detect differences in means or medians of two populations (for example, site and background populations).

- Ranked set sampling can be used in other sampling designs such as stratified random sampling and composite sampling.

- Ranked set sampling can be used to obtain more representative data for purposes other than estimating a mean by covering more of the target population. Such purposes include computing a confidence limit on the median of a population (Hettmansperger, 1995), testing for differences in the medians of two populations (Bohn and Wolfe, 1992, 1994), conducting simple tests to check for compliance with a fixed remediation concentration limit (Hettmansperger, 1995; Koti and Babu, 1996; Barabesi, 1998), estimating the slope and intercept of a straight line relationship (Muttlak, 1995), estimating the ratio of two variables (Samawi and Muttlak, 1996), and estimating the means of several populations in an experimental setting (Muttlak, 1996).

When the objective of sampling is to estimate the mean, consideration should be given to using ranked set sampling rather than simple random sampling when the cost of ranking potential sampling locations in the field is negligible or very low compared to the cost of laboratory measurements. Guidance on setting up a ranked set sampling design taking cost considerations into account, including ranking costs, is provided in Appendix 8-A.

## 8.4 LIMITATIONS

Before ranked set sampling is used, the costs of locating and ranking potential sampling locations in the field should be determined to make sure that ranked set sampling is cost-effective. Ranked set sampling can yield a more precise estimate of the population mean, the costs may be higher than if simple random sampling were used.

The precision of a mean that is computed using data obtained with ranked set sampling will be reduced if errors are made in ranking field locations. That is, the precision of the computed mean is maximized (i.e., the variance of the computed mean is minimized) when there are no errors in ranking field locations. However, even when professional judgment or on-site methods cannot rank field locations without error, ranked set sampling will perform as well as simple random sampling in estimating the mean for the same number of measurements.

In ranked set sampling, the field locations being compared (ranked) are supposed to be randomly located over the population. However, in practice, field locations within a set may be purposely clustered in close proximity to decrease the effort of taking screening measurements or to increase the accuracy of visually ranking the locations. In this case, the precision of the estimated mean obtained using ranked set sampling data may be reduced. To reduce or eliminate this decrease in

precision McIntyre (1952) suggests dividing the population into portions of equal size that have no well-defined gradients and then selecting an equal number of samples within each portion.

If ranked set sampling data are used to test hypotheses, the data computations may differ from the standard computations that would be performed if the data were obtained using simple random sampling. For example, suppose the Wilcoxon Rank Sum test will be used to test for differences in the medians of two populations and that the data are obtained using ranked set sampling. Then the data computations for the Wilcoxon Rank Sum test described in Bohn and Wolfe (1992, 1994) should be used rather than the standard computations [for example, see Section 18.2 of Gilbert (1987)] that would be used if the data had been obtained using simple random sampling. If ranked set sampling data will be used to conduct tests of hypotheses or to compute confidence intervals on means or other statistical parameters, guidance from a statistician familiar with ranked set sampling should be sought.

Finally, Appendix 8-A shows that the on-site measurements (for example, the XRF measurements in the above example) obtained for the ranking process are not used quantitatively when computing the estimated mean or the variance of the estimated mean. Hence, ranked set sampling does not make full use of the information content of the XRF measurements. One approach for making fuller use of on-site measurements is to use the "Double Sampling" design described in Section 9.1 of Gilbert (1987). In that design, the XRF measurements are used in combination with the lead measurements in a linear regression equation to estimate the mean. However, the Double Sampling design requires the XRF and lead measurements to be linearly related with a high correlation; ranked set sampling does not.

## 8.5 IMPLEMENTATION

### 8.5.1 How Do You Decide the Number of Samples for Laboratory Analysis Needed to Estimate the Mean?

Most methods in the statistical literature for determining the number of samples for estimating the mean were developed assuming that sampling locations are identified using simple random sampling rather than ranked set sampling. In general, ranked set sampling needs fewer samples than simple random sampling because ranked set sampling yields more information per set of measurements. This concept was illustrated in Table 8-1 for the normal distribution. Appendix 8-A provides a step-by-step process for determining the ranked set sampling sample size for estimating a mean.

Methods for computing the ranked set sampling sample size (number of samples for laboratory measurement) for other sampling objectives, such as testing hypotheses, are less well-developed and not yet available in the statistical literature. However, since ranked set sampling increases the performance of statistical procedures relative to what would be achieved if simple random sampling were used, the "*n*" calculated for simple random sampling should be adjusted to allow for a multiple of cycles (see the example in Appendix 8-A).

### 8.5.2 How Do You Decide Where in the Field to Collect Samples for Laboratory Analysis?

Locations at which samples for laboratory analysis will be collected are determined by the ranking process using professional judgment or on-site measurements. The use of ranked set sampling to determine the field locations is illustrated in Appendix 8-A for a balanced ranked set sampling design. In a balanced ranked set sampling design, the same number of locations are collected for each rank. For example, the simple ranked set sampling lead example given in Section 8.1 was a balanced design because the design needs an equal number of locations expected to have relatively low, medium, or high lead concentrations. A balanced ranked set sampling design should be used if the underlying distribution of the population is symmetric.

In an unbalanced ranked set sampling design, different numbers of locations expected to have relatively low, medium, or high concentrations are sampled. Environmental data are often asymmetric and skewed to the right; that is, with a few measurements that are substantially larger than the others. If the goal is to estimate the mean using ranked set sampling, McIntyre (1952) indicates the mean would be more precisely estimated if more locations expected to have relatively high concentrations were selected than locations expected to have relatively low or medium concentrations. This idea is discussed further by Patil et al. (1994). To illustrate an unbalanced ranked set sampling design, one could modify the lead example in Section 8.1 to collect a soil sample at twice as many locations expected to have relatively high lead concentrations as at locations expected to have relatively low or medium concentrations. When an unbalanced ranked set sampling design is used, the true mean of the population is estimated by computing a weighted mean, as described in Appendix 8-A, rather than the usual unweighted mean.

An appropriate unbalanced ranked set sampling design should increase the precision of the estimated mean of an asymmetric distribution. However, an inappropriate unbalanced ranked set sampling design for an asymmetric distribution can provide a less precise estimate of the mean than a balanced ranked set sampling design or a simple random sampling design. Kaur et al. (1995) established a method for developing an appropriate unbalanced ranked set sampling design for asymmetric distributions that are skewed to the right. This method is provided in Appendix 8-A.

## 8.6 EXAMPLES

### 8.6.1 Estimating Mean Plutonium Concentrations in Soil

Gilbert (1995) illustrates the use of ranked set sampling to obtain samples for estimating the mean plutonium (Pu) concentration in surface soil at some weapons testing areas on the Nevada Test Site. Pu concentrations in soil samples are typically measured in the laboratory, and measurement is quite expensive. However, at the weapons testing areas in Nevada**,** inexpensive field measurements of Americium-241 (denoted by $^{241}$Am) in surface soil can be obtained using an in-situ detector called the

FIDLER (Field Instrument for the Detection of Low Energy Radiation). Past studies had shown that in areas of high soil Pu concentrations, there is a relatively high correlation (about 0.7) between a FIDLER reading at a field location and a Pu measurement made on a 10-gram aliquot for a surface (0-5 centimeters) soil sample collected at that spot. Moreover, the cost of a Pu measurement in the laboratory is at least 10 times greater than the cost of obtaining a FIDLER reading. Hence, using Table 8-3 in Appendix 8-A, it appears that using ranked set sampling instead of simple random sampling to determine locations to collect soil samples for laboratory analysis should provide a more precisely estimated mean. Gilbert (1995) illustrates how to compute the mean and its variance using data from a balanced ranked set sampling design. It should be noted that, because the distribution of Pu measurements at the study areas is typically skewed to the right, an unbalanced ranked set sampling design might produce a more precise estimated mean than a balanced ranked set sampling design.

### 8.6.2 Estimating Mean Reid Vapor Pressure

Nussbaum and Sinha (1997) discuss a situation where ranked set sampling appears to have great potential for cost savings. Air pollution in large cities is currently being reduced through the use of reformulated gasoline. Reformulated gasoline was introduced because of regulations that limit the volatility of gasoline, as commonly measured by the Reid Vapor Pressure (RVP). Typically, RVP is measured on samples from gasoline stations obtained using simple random sampling. RVP can be measured in the laboratory or at the pump itself. Although laboratory measurement costs are not unduly expensive, it is expensive to ship samples to the laboratory. Hence, reducing the number of samples analyzed in the laboratory could result in a large costs savings without sacrificing the assessment of compliance with the volatility regulations.

One possible way to reduce the number of samples analyzed in the laboratory is to use ranked set sampling. Measurements of RVP taken at the pump might be used to rank samples using the ranked set sampling procedure to determine which samples should be taken to the laboratory for measurement. Suppose that (1) the correlation between field RVP and laboratory RVP measurements is sufficiently high so that the ranking was very accurate and that (2) it is several times more costly to transport and measure samples in the laboratory than it is to rank samples at the pump. In this case, the number of samples measured in the laboratory could be reduced by perhaps a factor of 2 or more without reducing the ability to determine when the volatility regulations are being violated. Nussbaum and Sinha (1997) present data that strongly suggest a very strong positive linear relationship between pump and laboratory measurements of RVP. This information may be used to justify the use of field RVP measurements to accurately rank the pump samples (see Table 8-2 in Appendix 8-A). Assuming no ranking errors, Table 8-2 shows that if the ratio of laboratory transportation and measuring costs to ranking costs (i.e., the cost of the field RVP measurement and ranking process) is greater than 6, then ranked set sampling can be expected to yield as precise an estimate of the mean RVP as what would be obtained using simple random sampling but at less cost.

### 8.6.3 Estimating Mean Pool Area in Streams

Mode et al. (1999) provided this example of a U.S. Department of Agriculture Forest Service data collection effort on Pacific Northwest streams as part of a large scale monitoring project. There was interest in assessing salmon production in streams. The size of salmon habitat, particularly pool area in streams, has been linked to salmon production. Obtaining pool area by accurately and precisely measuring length and width of stream pools is time consuming and labor intensive. However, visual estimates of pool area can be obtained at much less cost. Mode et al. (1999) found that ranked set sampling estimates of the mean pool area for 20 of 21 streams were more precise than estimates of the pool area that would be obtained by physically measuring pool areas selected using simple random sampling. They also found that for over 75% of the streams, it would be less costly to use ranked set sampling than simple random sampling to obtain the same precision in the estimated mean pool area when pool measuring costs were at least 11 times greater than the costs of visually assessing pool area.

## APPENDIX 8-A

## USING RANKED SET SAMPLING

### INTRODUCTION

This appendix provides guidance on how to develop a balanced or unbalanced ranked set sampling design and how to estimate the mean and the standard deviation of the mean based on the data obtained. Developing a ranked set sampling design for the purpose of estimating the mean of the population is a two step process:

Step 1. Determine if ranked set sampling is cost effective compared to simple random sampling. This step is accomplished by considering the costs and performance of professional judgment and inexpensive on-site methods for ranking field locations.

Step 2. If ranked set sampling is expected to be more cost effective than simple random sampling, then determine the number of samples for laboratory analysis needed to estimate the mean with the specified accuracy and confidence.

Details of how to implement Steps 1 and 2 are provided in this appendix along with the methods for computing the mean and its standard deviation.

### HOW DO YOU DECIDE IF RANKED SET SAMPLING IS MORE COST EFFECTIVE THAN SIMPLE RANDOM SAMPLING FOR ESTIMATING THE MEAN?

This section provides guidance on how to determine if ranked set sampling will be more cost effective than simple random sampling when the objective of sampling is to estimate the mean with a specified precision. Ranked set sampling is more cost effective than simple random sampling for estimating the mean if the cost of using professional judgment or on-site measurements to rank potential sampling locations is negligible (Patil et al., 1994). This conclusion stems from the fact that fewer samples for laboratory analysis are needed to estimate the mean with specified precision if ranked set sampling is used than if simple random sampling is used. Hence, laboratory measurement costs will be lower. However, ranking potential sampling locations in the field may be costly due to factors such as spending more hours in the field, locating and training an expert to subjectively rank field locations, and purchasing and using on-site field technologies. The basic question is whether the increased precision in the mean that can be obtained using ranked set sampling will compensate for the extra work and cost of ranking.

The effect of costs on the decision of whether to use ranked set sampling or simple random sampling can be approximated using Table 8-2. This table shows the approximate cost ratio (cost of a laboratory measurement divided by the cost of ranking a field location) that must be exceeded before ranked set sampling will be more cost effective than simple random sampling to estimate the mean with a desired level of precision. The cost ratio that must be exceeded depends on the set size, $m$ (number of locations sampled in each of the $r$ ranked set sampling cycles), and on the distribution of the population of laboratory measurements. Table 8-2 gives approximate cost ratios for normal measurements when there is different sizes of ranking error. Table 8-2 shows that for a given set size, the cost ratios that apply when there is substantial ranking error are almost double the ratios when there is no ranking error.

**Table 8-2. The Approximate Cost Ratio* for Estimating the Mean**

| Data Distribution | Degree of Ranking Error | Set Size m = 2 | Set Size m = 3 | Set Size m = 5 |
|---|---|---|---|---|
| Normal | None | 4 | 3.25 | 2.75 |
| Normal | Moderate | 5.5 | 5 | 4.5 |
| Normal | Substantial | 7.25 | 6.25 | 6.5 |

Constructed from Figure 3 in Mode et al. (1999).
*Cost of a laboratory measurement divided by the cost of ranking a field location.

Suppose that practical aspects of ranking in the field lead to using a relatively small set size of $m = 3$ and that prior studies at the site of interest indicate that laboratory measurements for the contaminant of interest are likely to be approximately normally distributed. Since the normal distribution is symmetric, a balanced ranked set sampling design will be used (a balanced design is defined in Section 8.5.2). If no errors are expected in ranking field locations, the ratio of laboratory measuring costs (per sample) to ranking cost (per field location) must be greater than approximately 3.25 in order for ranked set sampling to be more cost effective than simple random sampling; that is, for the total cost of ranked set sampling to be less than the total cost of simple random sampling to estimate the mean with a desired specified precision. If there is substantial ranking error and $m = 3$ is used, the cost ratio must be greater than 6.25 for ranked set sampling to be more cost effective than simple random sampling. However, if past studies indicate that the measurements are more likely to have a distribution that is skewed to the right, the cost ratios will have to be higher before ranked set sampling is efficient.

Note that the cost ratios in Table 8-2 were developed assuming that a balanced ranked set sampling design will be used. If the distribution of laboratory measurements is expected to be skewed to the right, then an unbalanced ranked set sampling design will be more efficient than a balanced ranked set sampling design.

The cost ratios in Table 8-2 can be used when field locations are ranked using either professional judgment or on-site measurements. Table 8-3 provides cost ratios from Figure 4 of Mode et al. (1999) for balanced ranked set sampling designs with set sizes *m* equal to 2, 4, 6, and 8 that are applicable when there is quantitative information on the correlation between the on-site measurement at a location and the measurement obtained in the laboratory for a sample collected at the field location. If the on-site measurement is a good predictor of the corresponding laboratory measurement, then the correlation between the two measurements will be close to 1 and no or very few ranking errors will occur. A correlation of exactly 1 implies no ranking errors. If the screening measurement has absolutely no ability to predict the value of the laboratory measurement, then the correlation will be zero.

**Table 8-3. Approximate Cost Ratio* for Estimating the Mean when On-site Measurements** Are Used to Rank Field Locations**

| Correlation (Degree of Ranking Error) | Set Size m = 2 | Set Size m = 4 | Set Size m = 6 | Set Size m = 8 |
|---|---|---|---|---|
| 1.0 (No ranking error) | 5 | 3 | 2 | 2 |
| 0.9 | 6 | 5 | 5 | 5 |
| 0.8 | 7 | 8 | 8 | 9 |
| 0.7 | 12 | 12 | 14 | 16 |

*Cost of a laboratory measurement divided by the cost of ranking a field location.
**Cost ratios are from Figure 4 of Mode et al. (1999) and were derived assuming the on-site measurements and the measurements in the laboratory have a bi-variate normal distribution.

If the correlation between the screening and laboratory measurements is close to 1, then the information gained by ranked set sampling via the ranking process increases appreciably compared to simple random sampling. Hence, the cost ratio need not be so large for ranked set sampling to be worth the extra effort and cost of ranking. For example if the correlation is 1, indicating no ranking errors, then the cost ratio can be as small as 2 or 3 for set sizes of $m = 4$ or larger. But ranking errors will occur if the correlation is 0.8 or smaller, and the additional information obtained using ranked set sampling will be reduced compared to simple random sampling. Consequently, the cost ratio that must be exceeded for ranked set sampling to be more cost effective than simple random sampling is relatively high (8 or more).

Tables 8-2 and 8-3 permit summary statements like the following (adapted from Mode et al., 1999): If the cost for a laboratory measurement is about six times that of a screening measurement or professional judgment determination, and given that past data sets have been fairly normally distributed, then ranked set sampling will be more cost effective than simple random sampling unless the chosen

ranking method will result in substantial ranking errors (Table 8-2) or is based on a on-site measurement that is not very highly correlated (Table 8-3).

It should be noted that the use of field measurements has advantages that can lower the cost of the overall project, such as by reducing the number of return trips to the field through using a dynamic work plan. Hence, on-site measurements can result in greater project cost savings than is apparent in a simple comparison of per sample costs as is done above.

## HOW DO YOU DETERMINE THE NUMBER OF SAMPLES FOR LABORATORY ANALYSIS TO ESTIMATE THE MEAN WHEN RANKED SET SAMPLING IS USED?

This section begins by defining and discussing the *relative precision* of ranked set sampling to simple random sampling. The relative precision is used in the process subsequently discussed for approximating the number of samples ("sample size") for laboratory analysis needed for balanced and unbalanced ranked set sampling designs.

### What is the *Relative Precision* of Ranked Set Sampling to Simple Random Sampling?

For a sample size *n*, the relative precision of ranked set sampling to simple random sampling is defined to be:

$$RP = Var(\overline{x}_{SRS}) / Var(\overline{x}_{RSS}) \qquad (8.1A)$$

where:

$Var(\overline{x}_{SRS})$ = variance of the estimated mean of the laboratory measurements if simple random sampling is used to select sampling locations, and

$Var(\overline{x}_{RSS})$ = variance of the estimated mean of the laboratory measurements if ranked set sampling is used to select the sampling locations.

Note from Equation (8.1A) that values of the relative precision greater than 1 imply that $Var(\overline{x}_{RSS})$ is less than $Var(\overline{x}_{SRS})$, in which case ranked set sampling should be considered for use instead of simple random sampling, assuming the applicable cost ratio in Table 8-2 or 8-3 is exceeded.

It is known (Patil et al., 1994) that the relative precision of ranked set sampling to simple random sampling is always equal to or greater than 1 when a balanced design is used, regardless of the shape of the distribution of the laboratory measurement data. This means that $Var(\overline{x}_{RSS})$ is always

expected to be less than Var( $\overline{x}_{SRS}$ ), a rather remarkable result. To be more specific, if a balanced ranked set sampling design is used, then:

$$1 \# RP \# (m + 1) / 2 \qquad (8.2A)$$

where *m* is the set size. For example, if *m* = 2, then the value of the relative precision is between 1 and 1.5, and if *m* = 3, then the relative precision is between 1 and 2. The particular value of the relative precision for any given study population depends on the distribution of the laboratory measurements. Given ranking can be achieved at little or no error, the upper bound of the relative precision, (*m*+1)/2, is achieved when the distribution of the measurements is rectangular. The relative precision lies between 1 and (*m*+1)/2 for all other distributions. The lower bound for the relative precision, 1, occurs when ranking is completely random, that is, when professional judgment or on-site measurements have no ability whatsoever to correctly rank field locations.

To use the sample size procedures given below, you need to specify a value for the relative precision. Patil et al. (1994) provide values of the relative precision for balanced ranked set sampling designs for normal, rectangular, beta, gamma, Weibull, exponential and several other distributions for set sizes, m, of 2, 3, 4, and 5. Patil et al. (1994) also provides values of the relative precision for balanced ranked set sampling designs for lognormal measurement distributions for set sizes, m, between 2 and 10. A portion of their relative precision values for the lognormal distribution are provided in Table 8-4. It should be noted that the relative precision values in Table 8-4 for lognormal distributions with CV = 0.10 are also appropriate for normal distributions. This occurs because a lognormal distribution with a very small CV value has a shape very similar to a normal distribution. As the CV becomes large, the lognormal distribution has a longer and longer tail extending to high data values.

The relative precisions in Table 1 of Patil et al.(1994) for gamma and Weibull (asymmetrical) distributions bracket a range of relative precision values that is similar to those for the lognormal distribution for the same set sizes, m. In practice, it is usually not known with high confidence whether the data should be modeled using a lognormal, gamma, Weibull, or some other right-skewed distribution, the relative precision values in Table 8-4 for the lognormal distribution are used here to approximate the number of samples (for laboratory analysis) needed to estimate the mean when a balanced ranked set sampling design is used.

**How Do You Determine the Number of Samples for Laboratory Analysis for Balanced Ranked Set Sampling Designs?**

The procedure for approximating the number of samples for laboratory analysis, *n*, needed for a balanced ranked set sampling design to estimate the population mean with specified precision and confidence is as follows:

**Table 8-4. Relative Precision (RP)* of Balanced Ranked Set Sampling to Simple Random Sampling for Lognormal Distributions**

| Set Size (m) | CV = 0.1** | CV = 0.3 | CV = 0.5 | CV = 0.8 |
|---|---|---|---|---|
| 2 | 1.5 | 1.4 | 1.4 | 1.3 |
| 3 | 1.9 | 1.8 | 1.7 | 1.5 |
| 4 | 2.3 | 2.2 | 2.0 | 1.8 |
| 5 | 2.7 | 2.6 | 2.3 | 2.0 |
| 6 | 3.1 | 2.9 | 2.6 | 2.2 |
| 7 | 3.6 | 3.3 | 2.8 | 2.4 |
| 8 | 3.9 | 3.6 | 3.1 | 2.5 |
| 9 | 4.3 | 3.9 | 3.3 | 2.7 |
| 10 | 4.7 | 4.3 | 3.6 | 2.9 |

* Values of relative precision are from Table 2 in Patil et al. (1994).

** CV = Coefficient of variation for the lognormal distribution, which is defined to be $CV = (\exp[\sigma^2] - 1)^{1/2}$, where $\sigma^2$ is the variance of the natural logarithms of the data.

Step 1: Use the DQO Process to determine the number of samples for laboratory analysis, $n_o$, needed to estimate the mean with specified accuracy and confidence if simple random sampling is used to determine the sampling locations. The method for determining $n_o$ is provided in Chapter 4.

Step 2 : Select a value of the set size, $m$. This value is usually based on practical constraints in ranking locations in the field using professional judgment or on-site measurements. It may be difficult to use professional judgment to accurately rank by eye more than 4 or 5 locations, which implies $m$ should not exceed 4 or 5. Other constraints that may affect the size of $m$ are time, staff, and cost considerations.

Step 3: Use the site conceptual model in conjunction with available data or information from prior studies or from new data collected at the site (from the same population) to select a value of the relative precision. Do this by first computing the estimated coefficient of variation (CV) of data collected previously from the same or very similar site using similar collection, handling, and measurement methods or by making use of probability plots or statistical techniques to determine if normality can be assumed (EPA, 2000b). Ideally, the number of

data (*N*) used to compute the CV should be at least 10. The estimated CV is computed as follows:

$$\text{Estimated CV} = s / \bar{x}$$

where:

$$\bar{x} = \sum_{i=1}^{N} x_i / N$$

$$s = [\sum_{i=1}^{N} (x_i - \bar{x})^2 / (N - 1)]^{1/2}$$

$x_i$ = the $i^{th}$ data value

and N = the number of data values used to compute the CV.

Use Table 8-4 with the computed value of the CV and the selected value of the set size m to determine the approximate value of relative precision (RP).

Step 4: Compute the number of replications (cycles), *r*, as follows:

$$r = ( n_o / m ) \times (1 / RP) \qquad (8.3A)$$

Step 5: Compute the total number of samples for laboratory analysis, *n*, that should be collected to estimate the mean:

$$n = r \times m$$

Note from Equation 8.3A that if RP = 1, then $r = n_o / m$ and hence $n = n_o$. Values of RP equal to 1 occur if the professional judgment or on-site measurements used have no ability whatsoever to correctly rank field locations, in which case ranked set sampling has no advantage over simple random sampling. In that case, the number of samples, n, needed by ranked set sampling is the same as that needed by simple random sampling. The only added cost has been the effort needed to select the unused sample locations. The factor 1/RP, which equals $Var[\bar{x}_{RSS}] / Var[\bar{x}_{SRS}]$ in Step 4, adjusts (decreases) the value of *r* to account for the fact that $Var(\bar{x}_{RSS}) < Var(\bar{x}_{SRS})$ whenever RP > 1.

Also, Table 8-4 shows that the relative precision is closer to 1 if the selected set size *m* is very small (for example, 2) and the CV is very large (indicating a highly skewed distribution). Hence, in this situation the number of samples, n, needed for a balanced ranked set sampling design to estimate the mean of a highly skewed distribution will be only slightly less than the number of samples needed by

simple random sampling. If the CV is large, consideration should be given to using an unbalanced ranked set sampling design as discussed later in this section.

**Example**

This example expands on the lead contamination example in Section 8.2 in the main text of this chapter. Suppose the goal is to estimate the mean concentration of lead in the surface soil of a residential property and that no major spatial patterns of lead concentrations are expected at the site. This suggests that simple random sampling may be considered for determining where soil samples should be collected for measurement of lead in the laboratory. (Stratified random sampling would have been considered if major spatial patterns existed and had been identified previously.)

However, suppose past studies concerning similar ranges of values expected to be found in this study had indicated that measurements of lead in soil obtained in the field using a hand-held x-ray fluorescence (XRF) in-situ detector have a correlation of approximately 0.9 with laboratory lead measurements made on soil samples collected at the measured field locations. This high correlation suggests that ranked set sampling might be used instead of simple random sampling in order to reduce the number of soil samples that would need to be measured in the laboratory. To determine if ranked set sampling would be more cost effective than simple random sampling, the cost of a laboratory measurement for lead was divided by the cost of ranking a field location using an XRF measurement to determine a measurement-to-ranking cost ratio. Suppose this cost ratio was found to be 10, such that ranking a field location is only one tenth as costly as a lead measurement in the laboratory. Table 8-3 shows that with a correlation of 0.9, the computed cost ratio (10) is greater than the tabled value of 6. Hence, it appears that ranked set sampling will indeed be more cost effective than simple random sampling and the number of samples that should be collected for laboratory analysis can be determined using the five-step process above.

Step 1: Determine the number of field samples for laboratory analysis, $n_o$, to estimate the true mean assuming that simple random sampling is used to identify the locations where samples will be collected.

It was determined using the method in Table 5-1 in Chapter 5 that using simple random sampling to determine field sampling locations would need a total of $n_o$ = 25 soil samples to estimate the mean lead concentration with 20% accuracy and 95% confidence.

Step 2: Select the set size, m.

The set size $m$ was selected to be $m = 5$. A larger value of $m$ was not used in order to limit time spent in the field to find and rank field locations where the XRF measurements would be taken.

Step 3: Determine the relative precision of ranked set sampling to simple random sampling.

Suppose that past studies on one or more similar residence properties had produced 50 soil samples that were collected using simple random sampling and that were handled, processed, and measured in the laboratory using the same or very similar methods as will be used in the present study. Displaying the data graphically using probability plots and histograms indicated that the data set was only slightly skewed to the right (to high lead concentrations). The CV of the population was estimated using the 50 measurements and was found to be 0.4. Hence, entering Table 8-4 with $m = 5$, the relative precision of a balanced ranked set sampling design was approximated to be about 2.45 (interpolating between RPs 2.6 and 2.3 for CV = 0.3 and 0.5, respectively).

Step 4: Determine r, the number of cycles of ranked set sampling.

The number of cycles, $r$, of ranked set sampling was computed as follows:

$$r = ( n_o / m ) \text{ x } (1 / \text{RP}) = (25/5) \text{ x } (1/2.45) = 2.04$$

that is rounded up to 3 to be conservative.

Step 5: Compute the total number of samples for laboratory analysis needed.

The total number of samples to be collected for the balanced ranked set sampling design is:

$$n = r \text{ x } m = 3 \text{ x } 5 = 15$$

as compared to $n_o = 25$ that would have been needed if simple random sampling is used.

The balanced ranked set sampling design is implemented by first identifying $m^2 = 5^2 = 25$ field locations using simple random sampling and then randomly dividing these 25 locations into 5 sets of size 5. The XRF detector ranks the five locations within the first set of five and a soil sample is collected at the location with the lowest XRF measurement. The second set of five locations is then

ranked using the XRF detector and a soil sample collected at the location with the second smallest XRF measurement in that set, and so on through the five sets of five locations to obtain five soil samples. Then that process is repeated $r = 3$ times to obtain a total of $r \times m = 3 \times 5 = 15$ soil samples that are measured for lead in the laboratory. (Note that this process needed a total of $m^2 r = 25 \times 3 = 75$ field locations to be measured by the XRF detector.) The (unweighted) arithmetic average of the 25 lead measurements is then computed to estimate the true mean lead concentration for the study area. The formulae for computing the mean and the variance of the estimate mean for both balanced ranked set sampling (as in this example) and unbalanced ranked set sampling designs are provided below.

**How Do You Compute the Mean and Variance of the Estimated Mean When Balanced Ranked Set Sampling Is Used?**

The true mean of the population is estimated by computing the arithmetic mean of the $n$ laboratory measurements obtained on the $n$ samples obtained using a balanced ranked set sampling. The formula is:

$$\overline{x}_{RSS,balanced} = (1/rm)\sum_{i=1}^{m}\sum_{j=1}^{r} x_{ij} \qquad (8.4A)$$

where

$r \times m = n$ = total number of samples obtained using a balanced ranked set sampling design

$x_{ij}$ = the measurement of the sample collected from the field location that had rank i that was collected in the $j^{th}$ cycle of sampling

The variance of the estimated mean $\overline{x}_{RSS,balanced}$ is computed as follows:

$$\text{Var}\left(\overline{x}_{RSS,balanced}\right) = \sum_{i=1}^{m}\sum_{j=1}^{r} (x_{ij} - \overline{x}_i)^2 / m^2 r(r-1) \qquad (8.5A)$$

where $x_{ij}$ was defined above and

$\overline{X}_i$ = the arithmetic mean of the $r$ laboratory measurements of the $r$ samples from field locations that had rank i collected during the $r$ cycles of sampling.

$$= (1/r)\sum_{j=1}^{r} x_{ij} \qquad (8.6A)$$

The standard deviation of $\overline{x}_{RSS,balanced}$ is the square root of Equation 8.5A.

**How Do You Determine the Number of Samples for Laboratory Analysis for Unbalanced Ranked Set Sampling Designs?**

The same two-step process is used to develop both a balanced and an unbalanced design: that is, to first determine if ranked set sampling is expected to be more cost effective than simple random sampling, and if so, then determine the number of samples for laboratory analysis to be collected. Although more research is needed to develop an optimal method to design an unbalanced ranked set sampling design, the "t-model" method developed by Kaur et al. (1995) appears to be a reasonable approach that should be satisfactory in practice. An unbalanced design should be considered if the distribution of the laboratory measurements is expected to be skewed to the right.

The "t-model" method consists of collecting $r$ samples for laboratory analysis for each of the $m$-1 smallest ranks ($m$ = set size) and $r$ x $t$ samples for the largest rank, where $t$ is some integer greater than 1. For example, if the set size is $m = 3$ and the number of cycles is $r = 5$, a *balanced* ranked set sampling design results in collecting a sample at each of five locations expected to have a relatively small value of the variable of interest (for example, lead), as well as at five locations expected to have a mid-value of lead and at five locations expected to have a relatively large value of lead. But for an *unbalanced* design, Kaur et al. (1995) suggest collecting a sample at 5 x $t$ locations (rather than five locations) expected to have relatively large values of lead, whereas the number of sample locations expected to have low or middle values of lead remain unchanged at 5. If the optimal value of $t$ is selected, then the relative precision of the unbalanced ranked set sampling design is greater than the relative precision of the balanced ranked set sampling design.

Optimum values of $t$ for various values of the CV for set sizes of $m = 2$, 3, 4 and 5 are plotted in Figure 6 of Kaur et al. (1995). The curves are essentially identical for these values of $m$. Their results are summarized in Table 8-5.

The total number of samples for laboratory analysis collected when the "t-model" method is used is computed as follows: $n = (m - 1 + t)\, r$, where $m$ is the prespecified set size, $r$ is the number of ranked set sampling cycles and $t$ is determined from Table 8-5. The formula for computing $r$ is given by Equation 8.3A, the same equation used for a balanced ranked set sampling design. However, the values of the relative precision used in Equation 8.3A will be too small if they are obtained from

**Table 8-5. Optimal Values of *t* for Determining the Number of Samples for Laboratory Analysis Needed for an Unbalanced Ranked Set Sampling Design**

| **CV** | 0.25 | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 |
|---|---|---|---|---|---|---|---|---|---|
| **t** | 1 | 2 | 3 | 5 | 6 | 7 | 8 | 9 | 10 |

Table 8-4 since those relative precision values apply to a *balanced* ranked set sampling design. In order to approximately correct for this bias, the values of relative precision in Table 8-4 should be multiplied by the correction factors in Table 8-6. The corrected relative precision values can then be used in Equation 8.3A to determine the approximate *r*. The correction factors in Table 8-6 are the approximate percent increase in the relative precision that occurs when an unbalanced "t-model" ranked set sampling design is used instead of a balanced design. For high values of CV ($i > 1.5$) the question of whether it is correct to be estimating mean quantities should be raised. For extremely skewed distributions, the choice of a difficult parameter, such as the median, should be considered.

**Table 8-6. Correction Factors* for Obtaining Relative Precision Values**

| **CV** | 0.1 | 0.3 | 0.5 | 0.8 | 1.3 |
|---|---|---|---|---|---|
| **Correction Factor** | 1.01 | 1.08 | 1.2 | 1.5 | 1.7 |

*Multiply the relative precision in Table 8-4 by these correction factors to obtain the approximate relative precision value to use in Equation 8-3A to determine the number of ranked set sampling cycles, r.

**Example**

This example illustrates the above description of how to determine the number of samples needed to estimate the mean when an unbalanced design is used. The lead contamination example that was used above will be used to illustrate a balanced ranked set sampling design. Recall that the correlation between laboratory lead measurements and the XRF in-situ detector readings of lead was approximately 0.9. Furthermore, the cost of a laboratory lead measurement divided by the cost of using XRF measurements to rank a field location was approximately 10. Hence, as the cost ratio 10 is larger than any of the cost ratios in Table 8-3 for any choice of *m* (set size) when the correlation is 0.9, ranked set sampling is expected to be more cost effective than simple random sampling for any value of *m* that might be used. Therefore, the five-step process used above for the balanced ranked set sampling design will be used to determine the number of samples that should be collected and measured for lead in the laboratory.

Step 1: Determine the number of field samples for laboratory analysis, $n_o$, needed to estimate the true mean assuming that simple random sampling is used to identify the locations where samples will be collected.

One approach for determining $n_o$ is to use the method given in Table 5-1 (Chapter 5). However, that method assumes that $n_o$ is large enough such that the estimated mean, $\overline{x}$, will have a normal distribution. However, if the distribution of the laboratory lead measurements is highly skewed (very asymmetrical), then the distribution of $\overline{x}$ may not be normal, in which case the

method in Table 5-1 (illustrated in Table 5-3) may provide a value of *n* that is too small. One way to mitigate this effect is to use a conservative (i.e., too large) value of the CV in the sample size formula, which will result in a larger value of $n_o$. Another approach, since the lead measurements are expected to be skewed (or else a balanced ranked set sampling design would be used rather than an unbalanced ranked set sampling design), is to use the method described in Section 3.6 of Perez and Lefante (1997) to determine the number of samples for estimating the mean of a *lognormal* distribution. Of course, if the distribution is skewed but not lognormal, then the number of samples obtained using their method may be too small or too large. The method in Table 5-1 is recommended for general use unless there is high confidence that the distribution is truly lognormal.

Suppose that past studies on similar residence properties had produced 50 lead measurements of the same type as will be obtained in the present study. Graphical displays of the data (probability plots, histograms and box plots) and statistical hypothesis tests for distribution shape indicated that the data were skewed to the right, but not necessarily lognormal. Also, the CV computed using the measurements was 0.7. But, this value was increased to 1.0 to help assure that the computed number of samples needed for laboratory analysis is not too small. Furthermore, during the DQO Process it was decided that the percent relative error of the estimated mean should be no more than 25%.

Hence, from Table 5-3 in Chapter 5, $n_o = 64$ samples are needed if simple random sampling is used to identify the locations where samples will be collected.

Step 2: Select the set size, *m*. Suppose the set size *m* was selected to be $m = 5$.

Step 3: Determine the value of the multiplicative factor *t*. For CV = 1.0, Table 8-5 shows that $t = 3$.

Step 4: Determine the value of r, the number of cycles of ranked set sampling for the $m-1 = 5-1 = 4$ smallest ranks:

$$r = ( n_o / m ) \times (1 / RP)$$

From steps 1 and 2 above, $n_o = 64$ and $m = 5$. Now, using Table 8-4 with $m = 5$ and CV = 1.0, the value of the relative precision obtained is 1.84 (using linear interpolation in the table). However, this value of relative precision needs

to be increased to correct for the (expected) skewness of the lead data set. The multiplicative correction factor is approximately 1.58, which is obtained by entering Table 8-6 with CV = 1.0 and using linear interpolation. Therefore, the correct value of the relative precision is:

$$RP = 1.84 \times 1.58 = 2.91$$

Therefore,

$$r = \frac{n_0}{m} x \frac{1}{RP} = \frac{64}{5} x \frac{1}{2.91} = 4.4, \text{round up to } 5.$$

Step 5. Compute the total number of samples for laboratory analysis needed if the unbalanced ranked set sampling design is used.

The total number of samples to be collected is:

$$n = (m + t - 1)\, r = (5 + 3 - 1)\, 5 = 35$$

and $r = 5$ soil samples will be collected for each of the $m - 1 = 5 - 1 = 4$ smallest ranks and $r \times t = 5 \times 3 = 15$ samples will be collected for the largest rank (rank 5).

The procedure for actually identifying which $n = 35$ field locations to sample is as follows:

1. Use simple random sampling to select $m + t - 1 = 5 + 3 - 1 = 7$ sets of $m = 5$ field locations.
2. Use the XRF in-situ detector to rank from 1 to 5 the five locations within each set.
3. For the first set of five locations, collect a sample at the location with rank 1.
4. For the second set, collect a sample at the location with rank 2.
5. For the third set, collect a sample at the location with rank 3.
6. For the fourth set, collect a sample at the location with rank 4.
7. For each of the fifth, sixth, and seventh sets, collect a sample at the location with rank 5.

The above procedure yields seven field samples for laboratory analysis. Now, repeat steps 1 through 7 a total of $r = 5$ times (cycles) to obtain the $n = 35$ samples needed. Note that $r \times t = 5 \times 3 = 15$ soil samples are collected at locations with rank 5 and $r = 5$ soil samples are collected for each of ranks 1, 2, 3, and 4, as needed by the "t-model" design.

**How Do You Compute the Mean and Variance of the Estimated Mean When Unbalanced Ranked Set Sampling Is Used?**

The true mean of the population is estimated by computing the weighted mean of the *n* laboratory measurements obtained on the *n* samples, where *ri* denotes the number in the $i^{th}$ cycle. The formula is:

$$\bar{x}_{RSS,unbalanced} = (1/m)\sum_{i=1}^{m}\left(\sum_{j=1}^{r_i} x_{ij} / r_i\right) \quad (8.7A)$$

The variance of the estimated mean, $\bar{x}_{RSS,unbalanced}$, is computed as follows:

$$\text{Var}(\bar{x}_{RSS,unbalanced}) = (1/m^2)\sum_{i=1}^{m}\left[\sum_{j=1}^{r_i}(x_{ij} - \bar{x}_i)^2 / r_i(r_i - 1)\right] \quad (8.8A)$$

Equations 8.7A and 8.8A simplify to equations 8.4A and 8.5A if $r_i = r_2 \ldots = r_m = r$, that is if a balanced ranked set sampling design is used. The standard deviation of $\bar{x}_{RSS,unbalanced}$ is the square root of Equation 8.8A.

It should be noted that the $7 \times 5^2 = 175$ XRF measurements obtained to rank the field locations are not quantitatively used in computing $\bar{x}_{RSS,unbalanced}$ or $\text{Var}(\bar{x}_{RSS,unbalanced})$. Hence, not all the information in the XRF measurements is used. For example, those measurements may provide information about the spatial distribution of lead over the study site. Also, consultation with a statistician familiar with ranked set sampling may identify an approach, perhaps similar to Double Sampling briefly discussed in Section 8.4, for estimating the mean and its variance using the XRF measurements in combination with the lead measurements.

# CHAPTER 9

# ADAPTIVE CLUSTER SAMPLING

## 9.1 OVERVIEW

Adaptive sampling designs are designs in which additional units or sites for observation are selected depending on the interpretation of observations made during the survey. Additional sampling is driven by the results observed from the initial sample. Several different types of approaches to this strategy are known as adaptive sampling designs; however, this chapter will only discuss adaptive cluster sampling.

Adaptive cluster sampling involves the selection of an initial probability-based sample. Typically, additional samples are selected for observation when a characteristic of interest is present in an initial unit or when the initial unit has an observed value meeting some prespecified condition (for example, when a critical threshold is exceeded). Choosing an adaptive cluster sampling design has two key elements: (1) choosing an initial sample of units and (2) choosing a rule or condition for determining adjacent units to be added to the sample.

Adaptive cluster sampling is useful in situations where the characteristic of interest is sparsely distributed but highly aggregated. Examples of such populations can be found in fisheries (shrimps clustering in large but scattered schools), mineral investigations (unevenly distributed ore concentrations), animal and plant populations (rare and endangered species), pollution concentrations and hot spot investigations, and epidemiology of rare diseases. Adaptive cluster sampling is most useful when quick turnaround of analytical results is possible (for example, with the use of field measurement technologies). Possible environmental applications of adaptive cluster sampling include soil remediation (investigating the extent of soil contamination), hazardous waste site characterizations, surveying Brownfields, and determining the extent of occurrence of effects of an airborne source of pollutant on nearby flora and fauna.

Note that adaptive cluster sampling is similar in some ways to the kind of "oversampling" done in many geostatistical studies. Reasonably unbiased estimates of the site mean can be garnered via either a) declustering techniques, b) polygons of influence, or c) kriging. Kriging also allows an estimate of the standard error of the mean once the pattern of spatial covariance has been modeled.

## 9.2 APPLICATION

Consider the following scenario for a contamination study. In most places sampled, contamination is light or negligible, but a few scattered pockets of high contamination are encountered. There are two questions of interest. First, what is the average level of contamination for the whole

area? Second, where are the hot spots located? Using the traditional statistical approach, a random or systematic sample of sites or units would be selected and the contaminant measured at each selected site. The average of these measurements provides an unbiased estimate of the population average. The individual observations can be used to create a contour map to locate peaks of contamination. However, with this pattern of contamination, the traditional statistical approach has problems. If contamination is negligible over most of the area, the majority of the measurements will be zero or have levels that are nondetectable. Further, random sampling may miss most of the pockets of higher concentration. Thus, even though the sample average is still an unbiased estimator of the population mean, it will be less precise than an unbiased estimator that takes into account the unevenness in the distribution of the contaminant over the entire area. Furthermore, the contour map from a simple random sample design may not be as accurate in the areas of higher contamination levels because the areas are not well-represented in the sample.

Adaptive cluster sampling could provide a better approach in situations similar to the one described above. For populations where the characteristic of interest is sparsely distributed but highly aggregated, adaptive cluster sampling can produce substantial gains in precision (i.e., lower variability) over traditional designs using the same sample sizes.

## 9.3 BENEFITS

Adaptive cluster sampling has several benefits. First, unlike traditional designs which focus only on one objective, it simultaneously addresses the objective of estimating the mean concentration and the objective of determining the extent of contamination. Adaptive cluster sampling concentrates resources in areas of greater interest. In a hot spot investigation, for instance, interest is on areas with high levels of contamination. Adaptive cluster sampling directs selection of additional sampling units to these high contamination areas, provided that the initial sample "hits" the areas of interest.

In addition, field technologies used in adaptive cluster sampling can provide quick turnaround time on test results and allow fewer sampling events. Finally, additional characteristics can be observed, adding to the overall usefulness of the study. For instance, in studies on the presence or absence of rare animal populations, measurements on size, weight, etc. can be made on the animals that are found.

## 9.4 LIMITATIONS

The iterative nature of adaptive cluster sampling introduces some limitations. With adaptive cluster sampling, the process of sampling, testing, resampling and testing may take considerable time. If quick and inexpensive field measurements are not readily available, the total sampling costs could quickly grow large. Because the sampling process stops only when no more units are found to have the characteristic of interest, the final overall sample size is an unknown quantity. This feature makes the

total cost also an unknown quantity. Although it is possible to budget for the sampling process using expected total cost, the expected total cost also depends strongly on the validity of the assumption that the characteristic of interest is not widely spread. Consider a contamination investigation where only a few small areas of high contamination are assumed. Suppose this assumption is not valid; that is, the contamination is more widespread, almost throughout the entire study area. The initial sample has a high probability of "hitting" a contaminated area. Because the contaminated areas are widespread, the follow-up sample size will be larger, so the total sample size will be closer to the number of sampling units in the population, resulting in a much higher total cost.

The statistical theory and analytical methodology pertaining to adaptive cluster sampling is currently limited to estimating means and variances. The sample mean and sample variance are unbiased estimators of the population mean and variance only if these are obtained from the initial probability-based sample. Appendix 9-A discusses some unbiased estimators of the mean and variance using the entire adaptive cluster sample. Current statistical studies are being made to obtain readily usable inferential tools (confidence intervals, hypothesis testing, etc.) that have been modified for adaptive cluster sampling.

Table 9-1 summarizes the main features between simple random sampling, grid sampling and adaptive cluster sampling with either an initial simple random sample or an initial grid sample.

**Table 9-1. Comparison of Designs**

| Feature | Conventional Sampling | | Adaptive Cluster Sampling | |
|---|---|---|---|---|
| | Simple Random Sample | Grid Sample | With Initial Simple Random Sample | With Initial Grid Sample |
| Unbiased estimators for mean and variance? | Yes | Yes | Yes | Yes |
| Confidence limits/hypothesis tests? | Yes | Yes † | Yes* | Yes*† |
| Quantifiable decision error rates? | Yes | Yes | Yes* | Yes* |
| Hot spot detection probabilities? | No | Yes | No | Yes* |
| Extent of detected hot spots? | No | No | Yes | Yes |
| Sample size computations feasible? | Yes | Yes | No | No |
| Sampling cost prediction feasible? | Yes | Yes | No | No |

*Only based on initial sample size

†Given the validity conditions discussed in Sections 7.3 and 7.4

## 9.5 IMPLEMENTATION

Adaptive cluster sampling design is implemented using the following basic elements:

- Selecting the initial probability-based sample,
- Specifying a rule or criterion for performing additional sampling, and
- Defining the neighborhood of a sampling unit.

To develop an adaptive sampling design, a grid is placed over a geographical area of interest (target population), in which each grid square is a potential (primary) sampling unit. This is illustrated in Figure 9-1(a). Shaded areas on the figure indicate the unknown areas of concern; for instance, areas of elevated contaminant levels. This example has three regions of contamination. The 10 darkened squares in the figure represent a randomly selected set of 10 sampling units constituting the initial sample. Selection of the initial sample design is further discussed in Section 9.6. Whenever a sampled unit is found to exhibit the characteristic of interest — that is, the unit intersects any part of the shaded areas — neighboring sampling units are also sampled using a consistent pattern. An example follow-up sampling pattern is shown in Figure 9-2, where the xs indicate the neighboring sampling units to be sampled. The follow-up sampling pattern is called the *neighborhood* of a sampling unit. For this follow-up sample, the five grid units in the figure make up the neighborhood of the initially sampled unit. In Figure 9-1(a), three initial sampling units exhibit the characteristic of interest. The units adjacent to these three initial units are sampled next, as shown in Figure 9-1(b). Some of these sampled adjacent units also exhibit the characteristic of interest, so the units adjacent to these are sampled next, as shown in Figure 9-1(c). Figures 9-1(d) to (f) show subsequent sampling until no more sampled units exhibit the characteristic of interest. Figure 9-3 shows the initial random sample and the final sample. Note that the final sample covers two of the three regions of interest. If at least one of the initial units had intersected the third area, it would also have been covered by a cluster of observed units.

In order to estimate the mean and variance, care must be taken as not all the samples are truly random. Special formulae in Appendix 9-A must be used. The final sample consists of *clusters* of selected (observed) units around the initial observed units. Each cluster is bounded by a set of observed units that do not exhibit the characteristic of interest. These are called *edge units*. An initial observed unit that does not exhibit the characteristic of interest is also considered a cluster of size one. A cluster without its edge units is called a *network*. Any observed unit, including an edge unit or an initial observed unit, that does not exhibit the characteristic of interest is a network of size one. Hence, the final sample can be partitioned into non-overlapping networks. In the final sample in Figure 9-3, there are 2 networks each with more than 1 unit and 40 networks of size 1 (33 edge units and 7 initial observed units). These definitions are important in understanding the estimators for statistical parameters like the mean and variance discussed in Appendix 9-A.

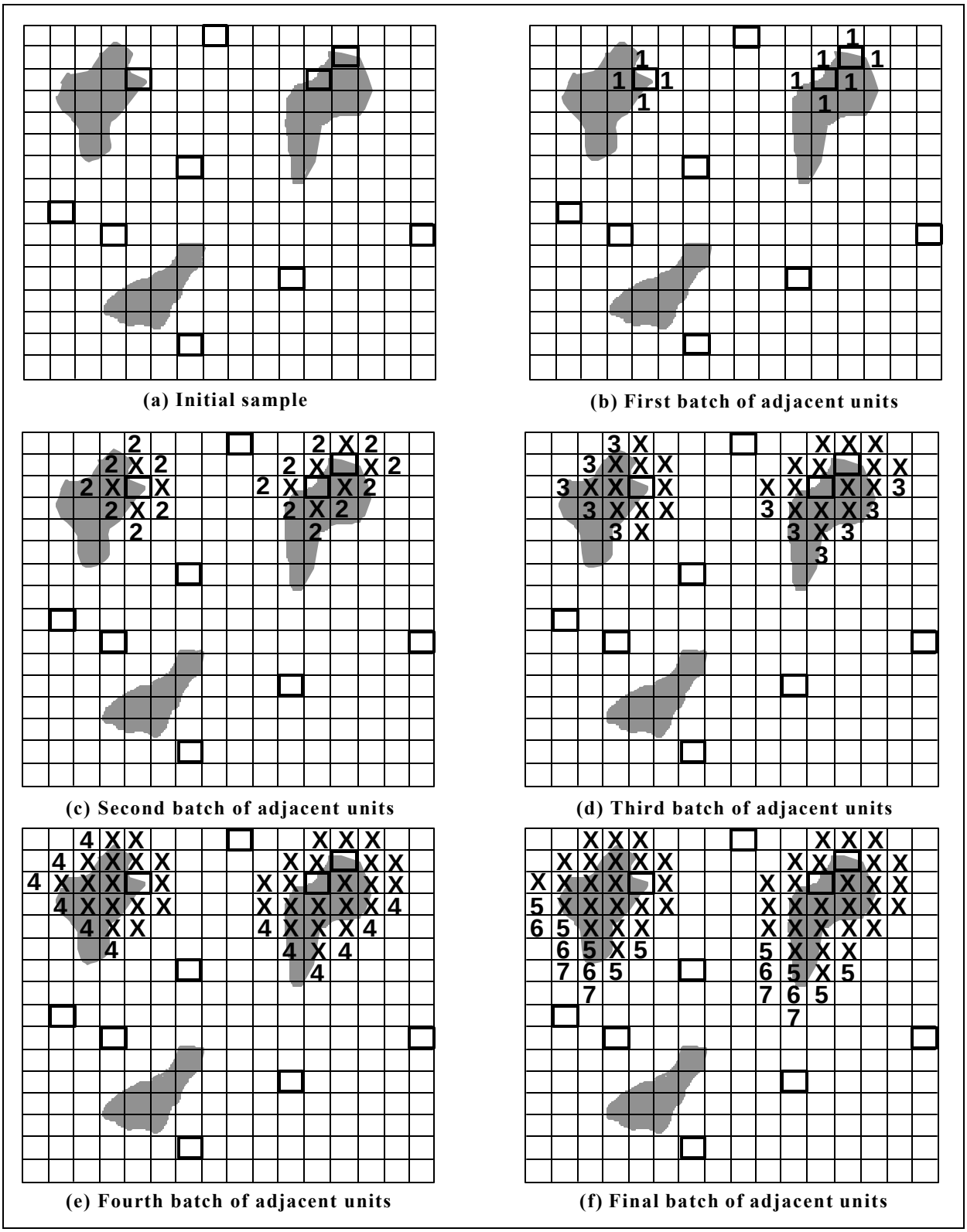


**Figure 9-1. Population Grid with Initial and Follow-up Samples and Areas of Interest**

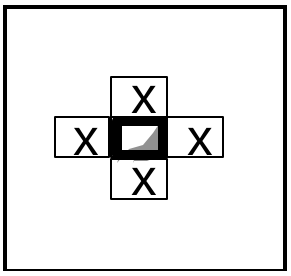

**Figure 9-2. Follow-up Sampling Pattern**

Note that using this design, one of the contaminated areas is missed in the sampling. This is a risk of

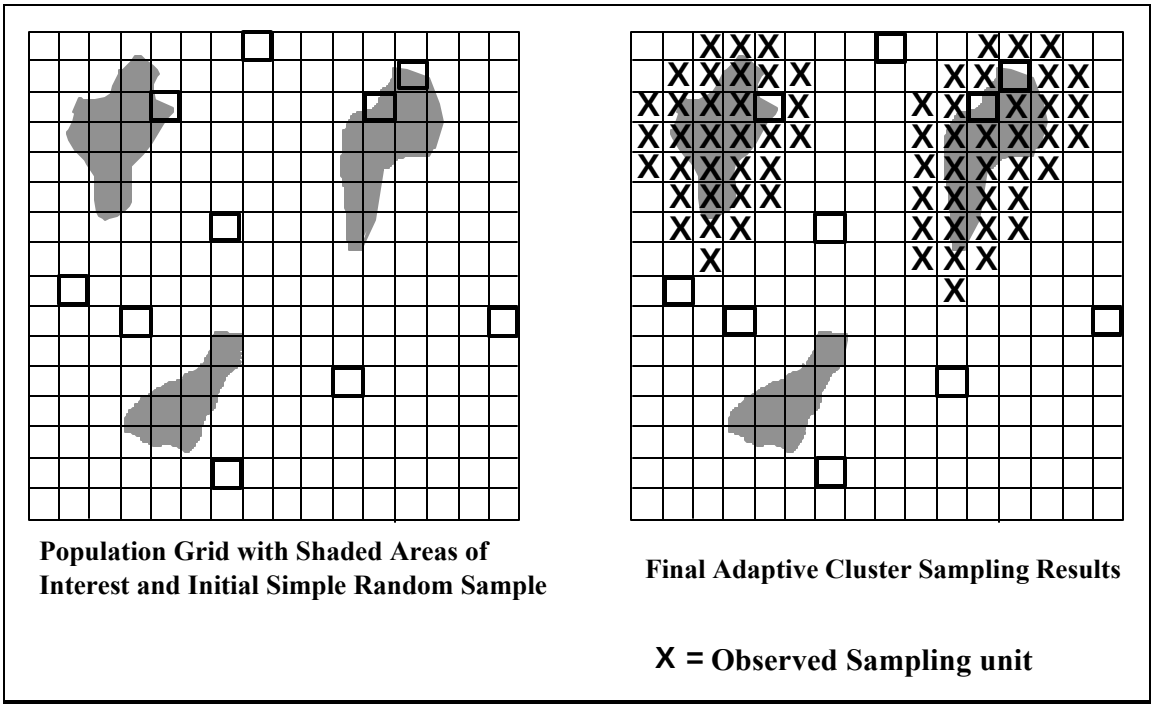


**Figure 9-3. Comparison of Initial Sample with Final Sample**

using adaptive sampling design. Although Adaptive Sampling may result in a large number of samples, it does outline the extent of the areas of interest.

## 9.6 RELATIONSHIP TO OTHER SAMPLING DESIGNS

The initial sample may be obtained using a number of traditional sampling designs. The choice of an initial sampling design is based on the available information about the distribution of the characteristic of interest: possible locations of aggregations or clusters, patterns of contamination, direction of contamination. If little is known of the extent or distribution of the characteristic of interest over the study region, an initial simple random sample may be useful. If prior information is available,

then a stratified sampling or grid sampling approach can be utilized. Section 7.5.2 discusses the use of grid sampling for finding hot spots, and Section 7.5.1 discusses the availability of software for determining optimal grid spacing.

An alternative scheme uses primary and secondary sampling units. Suppose the study area in Figure 9-1(a) is divided into vertical rectangular strips, each one square wide. The strips are the primary sampling units, each of which consists of secondary sampling units (squares). An initial random sample of strips is obtained and a random subset of squares (secondary sampling units obtained). If any one of the secondary units within a sampled strip is found to have the characteristic of interest, then the neighborhood of that secondary unit is sampled (the neighborhood is defined as in Figure 9-2). This scheme is particularly useful when sampling large areas. Section 4.7 of Seber and Thompson (1996) provide illustrations of these alternative schemes that use strips as the primary sampling unit.

## 9.7 EXAMPLE

Overflow from an impoundment containing nuclear liquid waste historically flowed into an adjacent field (Figure 9-4). In the field, the flow separated into multiple distinct channels before discharging into a river. The outflow has been shut off for 10 years, the field has been paved into a parking lot, and a new building has been proposed for that parking lot. There is no available information indicating the former locations of the flow channels (for example, no aerial photographs or surveys). The contamination distribution needs to be characterized to evaluate the potential for contamination to migrate into the adjacent river. Construction of a new building could expose workers and future building inhabitants to contamination. Considering that the contamination is likely to be clustered within the former flow channels and that little prior information is known about the specific locations of the channels (the area cannot easily be stratified), adaptive cluster sampling is an ideal sampling design for this situation. A grid would be established across the parking lot. An initial random sample would be collected, and wherever the concentration or radioactivity exceeds threshold values, neighboring locations would be sampled. Neighboring locations would be sampled in an iterative process until the entire distribution of contamination in the field is characterized.

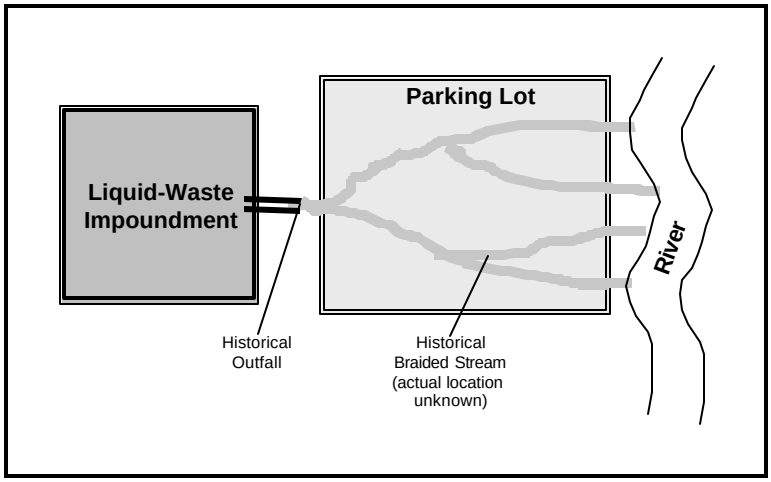


**Figure 9-4. Illustration of an Ideal Situation for Adaptive Cluster Sampling**

The initial sample could either be a simple random sample, a grid sample, or a strip sample with the strips oriented horizontally from the impoundment towards the river.

To obtain an initial strip sample, first divide the study region into horizontal strips of equal length and width. The region of interest here is the parking lot. Next, divide each strip into smaller, equal-sized areas. Select a simple random sample of strips and a further random sample of areas within strips. If any sampled area shows an exceedance over threshold, neighboring areas are investigated. Although this initial sample design stands the highest chance of capturing most of the contaminated areas, it also has the potential for a very large final sample size if the contamination is really not clustered around the former flow channels as assumed. If the contamination is more widespread, an initial simple random sample or grid sample may lead to a more cost-efficient final sample. Numerical examples in Appendix 9-A show how to calculate estimates of the mean and variance when using an adaptive cluster sampling design with an initial simple random sample.

## APPENDIX 9-A

## ESTIMATORS OF MEAN AND VARIANCE FOR ADAPTIVE CLUSTER SAMPLING WITH AN INITIAL SIMPLE RANDOM SAMPLE

[*Note to reader*: The following sections are intended for those with a strong statistics background. Software (Visual Sampling Plan) will be constructed to implement these designs thus alleviating the need for the algebra.]

Select a simple random sample of size $n_1$ from a population of $N$ units (for example, grid units or strip units). Define and determine the neighborhood of each unit. An example of a neighborhood would be an initial unit and the immediately adjacent units forming a cross (see Figure 9-2). For each unit, $I$, in the initial sample, determine whether or not the observed characteristic of interest satisfies a specified condition about a critical value $C$. For example, the observed characteristic of interest could be the measured amount of a certain contaminant observed in a sample unit, and the criterion could be exceedance of a critical value $C$ $(y>C)$. If so, sample all the units in the neighborhood of unit $I$. If any of the units in the neighborhood of unit $I$ satisfy condition $C$, then sample the neighborhood(s) of these units as well. Continue sampling neighborhood units until no more units satisfy the condition. The clusters of units in the sample are bounded by edge units, or units that do not satisfy the condition but are either included in the initial sample or are in the sampled neighborhoods in the follow-up sample. The units in each cluster that are not edge units form a network. Any observed unit, including edge units, that does not satisfy $C$ is considered to be a network of size one. This sampling design partitions the $N$ population units into distinct and disjoint networks.

Note that the method used for obtaining physical samples for analysis within each sample unit depends on the type of application. In particular, for environmental applications, study objectives and decision rules would determine if a single soil grab sample within each unit is adequate or whether or not a composite sample, obtained by combining soil from throughout the unit, should be used.

The usual sample average and sample variance (from a simple random sample) are going to be biased when calculated using the entire final sample. If only the initial sample is used for estimating the mean and variance, unbiased estimators based on the initial sample design can be obtained. Unbiased estimators of the mean and variance based on the final sample exist. However, these estimators involve more complex calculations than the estimators obtained from a simple random sample. The availability of computer software would greatly assist with these calculations and will be a separate document to be published in the near future.

Thompson (1990) has investigated unbiased estimators of the mean and variance based on the final sample. He developed modifications of the Horvitz-Thompson (Horvitz and Thompson, 1952)

and Hansen-Hurwitz (Hansen and Hurwitz, 1943) estimators. For an adaptive cluster sample with an initial simple random sample, the modified Horvitz-Thompson form of the estimators are:

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{\kappa}\frac{y_k^*}{\alpha_k} \tag{9A-1}$$

and

$$\hat{\text{var}}(\hat{\mu}) = \frac{1}{N^2}\left[\sum_{j=1}^{\kappa}\sum_{k=1}^{\kappa}\frac{y_j^* y_k^*}{\alpha_{jk}}\left(\frac{\alpha_{jk}}{\alpha_j\alpha_k} - 1\right)\right] \tag{9A-2}$$

where: $y_k^*$ = sum of the values of the character of interest, $y$, for the $k^{th}$ network in the sample
$N$ = number of units in the population
$\kappa$ = number of distinct networks in the sample
$a_k$ = probability that the initial sample intersects the $k^{th}$ network
$a_{jk}$ = probability that the initial sample intersects both the $j^{th}$ and the $k^{th}$ networks

Units in the initial sample that do not satisfy the condition $C$ are included in the calculation as networks of size one, but edge units are excluded.

If there are $x_k$ units in the $k^{\text{th}}$ network, then the intersection probabilities $\alpha_k$ and $\alpha_{jk}$ are calculated using combinatorial formulae as follows:

$$\alpha_k = 1 - \left[\frac{\binom{N - x_k}{n_1}}{\binom{N}{n_1}}\right] \tag{9A-3}$$

$$\alpha_{jk} = 1 - \left[\binom{N - x_j}{n_1} + \binom{N - x_k}{n_1} + \binom{N - x_j - x_k}{n_1}\right] \Big/ \left[\binom{N}{n_1}\right] \tag{9A-4}$$

where $a_{jj} = a_j$

**Example**

Consider the adaptive cluster sample shown in Figure 9-3. There are $N$=256 grid units in the population and $n_1$=10 units in the initial sample. One initial sample unit on the upper left area of the

study region intersected a network of $x_1$=18 units. Let this be network $A_1$. Two other initial sample units on the upper right area of the study region intersected a network ($A_2$) of $x_2$=19 units. The remaining seven initial sample units form networks of size one ($A_3$, $A_4$,...,$A_9$). Hence, there are $k$=9 distinct networks, with $x_1$=18, $x_2$=19, $x_3 = x_4 = ... = x_9 = 1$ units, respectively. The intersection probability for network $A_1$ is:

$$\alpha_1 = 1 - \left[ \binom{256-18}{10} \Big/ \binom{256}{10} \right] = 1 - \left[ \left( \frac{238!}{10!228!} \right) \Big/ \frac{256!}{10!246!} \right] = 0.5241791$$

while the intersection probability for network $A_2$ is:

$$\alpha_2 = 1 - \left[ \binom{256-19}{10} \Big/ \binom{256}{10} \right] = 1 - \left[ \left( \frac{237!}{10!227!} \right) \Big/ \frac{256!}{10!246!} \right] = 0.5441714$$

For the remaining networks ($A_3$, $A_4$,...,$A_9$) the intersection probability is:

$$\alpha_k = 1 - \left[ \binom{256-1}{10} \Big/ \binom{256}{10} \right] = 1 - \left[ \left( \frac{255!}{10!245!} \right) \Big/ \frac{256!}{10!246!} \right] = 0.0390625$$

Hence, the estimate of the mean using the Horvitz-Thompson estimator is:

$$\hat{\mu} = \frac{1}{256} \left[ \frac{y_1^*}{0.5241791} + \frac{y_2^*}{0.5441714} + \frac{y_3^* + y_4^* + ... + y_9^*}{0.0390625} \right]$$

where $y_1^*$ is the sum of the 18 observations from network $A_1$, $y_2^*$ is the sum of the 19 observations from network $A_2$, and $y_4^*$ , $y_9^*$ , ... $y_9^*$ are the single observations from the networks of size one. To compute an estimate of the variance, the joint intersection probabilities are needed:

$$\begin{aligned}
\alpha_{12} &= \alpha_{21} \\
&= 1 - \left[ \binom{256-18}{10} + \binom{256-19}{10} + \binom{256-18-19}{10} \right] \Big/ \left[ \binom{256}{10} \right] \\
&= 1 - \left[ \left( \frac{238!}{10!228!} \right) + \left( \frac{237!}{10!227!} \right) + \left( \frac{219!}{10!209!} \right) \right] \Big/ \left( \frac{256!}{10!246!} \right) \\
&= 0.2719547
\end{aligned}$$

$$\alpha_{jk} = \alpha_{kj} \text{ for } j = 3,4,\ldots,9 \text{ and } k = 3,4,\ldots,9, j \neq k$$

$$= 1 - \left[\binom{256-1}{10} + \binom{256-1}{10} + \binom{256-2}{10}\right] \Big/ \left[\binom{256}{10}\right]$$

$$= 1 - \left[\left(\frac{255!}{10!245!}\right) + \left(\frac{255!}{10!245!}\right) + \left(\frac{254!}{10!244!}\right)\right] \Big/ \left(\frac{256!}{10!246!}\right)$$

$$= 0.0013786$$

$$\alpha_{1j} = \alpha_{j1} \text{ for } j = 3,4,\ldots,9$$

$$= 1 - \left[\binom{256-18}{10} + \binom{256-1}{10} + \binom{256-19}{10}\right] \Big/ \left[\binom{256}{10}\right]$$

$$= 1 - \left[\left(\frac{238!}{10!228!}\right) + \left(\frac{255!}{10!245!}\right) + \left(\frac{237!}{10!227!}\right)\right] \Big/ \left(\frac{256!}{10!246!}\right)$$

$$= 0.0190701$$

$$\alpha_{2j} = \alpha_{j2} \text{ for } j = 3,4,\ldots,9$$

$$= 1 - \left[\binom{256-19}{10} + \binom{256-1}{10} + \binom{256-20}{10}\right] \Big/ \left[\binom{256}{10}\right]$$

$$= 1 - \left[\left(\frac{237!}{10!227!}\right) + \left(\frac{255!}{10!245!}\right) + \left(\frac{236!}{10!226!}\right)\right] \Big/ \left(\frac{256!}{10!246!}\right)$$

$$= 0.0198292$$

Then,

$$\hat{\text{var}}(\hat{\mu}) = \frac{1}{256^2}\left[\sum_{j=1}^{9}\sum_{k=1}^{9} \frac{y_j^* y_k^*}{\alpha_{jk}}\left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1\right)\right]$$

$$= \frac{1}{256^2}\left[\sum_{j=1}^{9} \frac{\left(y_j^*\right)^2}{\alpha_j}\left(\frac{1}{\alpha_j} - 1\right) + 2\sum_{j=1}^{8}\sum_{k=j+1}^{9} \frac{y_j^* y_k^*}{\alpha_{jk}}\left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1\right)\right]$$

$$= \frac{1}{256^2}\left[\frac{\left(y_1^*\right)^2}{\alpha_1}\left(\frac{1}{\alpha_1} - 1\right) + \ldots + \frac{\left(y_9^*\right)^2}{\alpha_9}\left(\frac{1}{\alpha_9} - 1\right)\right] + \frac{2}{256^2}\left[\frac{y_1^* y_2^*}{\alpha_{12}}\left(\frac{\alpha_{12}}{\alpha_1 \alpha_2} - 1\right) + \ldots + \frac{y_8^* y_9^*}{\alpha_{89}}\left(\frac{\alpha_{89}}{\alpha_8 \alpha_9} - 1\right)\right]$$

$$= \frac{1}{256^2}\left[\frac{\left(y_1^*\right)^2}{0.524179}\left(\frac{1}{0.524179} - 1\right) + \ldots + \frac{\left(y_9^*\right)^2}{0.0390625}\left(\frac{1}{0.0390625} - 1\right)\right]$$

$$+ \frac{2}{256^2}\left[\frac{y_1^* y_2^*}{0.2719547}\left(\frac{0.2719547}{(0.524179)(0.5441714)} - 1\right) + \ldots + \frac{y_8^* y_9^*}{0.0013786}\left(\frac{0.0013786}{(0.0390625)(0.0390625)} - 1\right)\right]$$

The second type of estimator is a modified Hansen-Hurwitz estimator and is based on the numbers of initial intersections. For an initial simple random sample, the estimators have the form:

$$\tilde{\mu} = \frac{1}{n_1}\sum_{i=1}^{N}\frac{y_i f_i}{m_i} = \frac{1}{n_1}\sum_{i=1}^{n_1} w_i = \overline{w} \qquad (9A\text{-}5)$$

and

$$\hat{var}(\tilde{\mu}) = \frac{N - n_1}{N n_1 (n_1 - 1)}\sum_{i=1}^{n_1}(w_i - \tilde{\mu})^2 \qquad (9A\text{-}6)$$

where: $y_i$ = value of the character of interest, *y,* for the $i^{th}$ unit
$n_1$ = number of units in the initial sample
N = number of units in the sample
$f_i$ = number of units in the initial sample which intersect network $A_i$ that includes unit *I*
$m_i$ = number of observations in the network $A_i$ that includes unit *I*
$w_i$ = $\frac{1}{m_i}\sum_{j\in A_i} y_j$ = mean of the $m_i$ observations in the network $A_i$ that includes unit *I*

**Example**

Consider again the adaptive cluster sample shown in Figure 9-3. For this example, N=256 and $n_1$=10. There is one initial sample unit in network $A_1$, two in network $A_2$ , and one each in networks $A_3$, $A_4$...$A_9$. Hence,

$$w_1 = 1/18\ y_1^*\ .\ w_2 = 1/19\ y_2^*\ .\ \text{and}\ w_j = y_j^*$$

for j=3, 4, ..., 9. As in the previous example, $y_j^*$ represents the sum of the observations in network $A_j$. The modified Hansen-Hurwitz estimators of the mean and variance are given by:

$$\tilde{\mu} = \frac{1}{10}\left[\frac{1}{18}y_1^* + \frac{1}{19}y_2^* + (y_3^* + \ldots + y_9^*)\right]$$

$$\hat{var}(\tilde{\mu}) = \frac{(256-10)}{256(10)(10-1)}\left[\left(\frac{1}{18}y_1^* - \tilde{\mu}\right)^2 + \left(\frac{1}{19}y_2^* - \tilde{\mu}\right)^2 + (y_3^* - \tilde{\mu})^2 + \ldots + (y_9^* - \tilde{\mu})^2\right]$$

Both estimators of the mean, $\hat{\mu}$ and $\tilde{\mu}$ are unbiased for the population mean :. The estimators of the variances, $\hat{var}(\hat{\mu})$ and $\hat{var}(\tilde{\mu})$ are also unbiased for $var(\hat{\mu})$ and $var(\tilde{\mu})$, respectively.

However, $\mathrm{var}(\tilde{\mu})$ tends to be slightly higher than $\mathrm{var}(\hat{\mu})$ (Christman, 2000). More examples of calculations for these estimators are given in Section 4.6 of Thompson and Seber (1996).

The relative efficiency of adaptive cluster sampling versus conventional sample designs can be measured by the ratio of the variances of the mean estimators from the designs being compared. Section 4.6 of Thompson and Seber (1996) discuss several factors that can increase the efficiency of adaptive cluster sampling designs (using the Hansen-Hurwitz estimator ($\hat{\mu}$):

- When within-network variability is a high proportion of total population variance, indicating clustered or aggregated populations (according to the character of interest).

- When there is a high degree of geographic rarity of the population, that is, when the number of units is large relative to the number of units satisfying the condition *C*, and the study region is large relative to the area where the contamination levels are high.

- When the expected final sample size is not much larger than the initial sample size (i.e., when the units satisfying the condition are clustered together in few clusters, and the units not satisfying the condition but included in the sample are also few in number).

- When units can be observed in clusters, which is less costly than observing the same number of units scattered at random throughout the region.

- When units observed do not satisfy the condition, which is less costly than observing units that satisfy the condition.

- When an easy-to-observe auxiliary variable is used to determine additional sampling; this can cut costs by eliminating the need to measure edge units.

Christman (1997) showed that the efficiency of adaptive cluster sampling relative to simple random sampling without replacement also depends on the choice of the condition C *(y>c)* and on the choice of neighborhood. As *c* increases, the within-network variance decreases, and the estimator becomes less efficient (see item 1 above). Also, using a neighborhood structure that does not consider the likely shape of the clusters of rare units may decrease efficiency. For instance, if the rare units tend to be physically adjacent, a neighborhood structure that includes physically adjacent units will tend to be more efficient than a neighborhood structure that does not.

## COST MODEL

It is difficult to derived advanced cost estimates for adaptive cluster sampling since sample sizes are random quantities. In some applications, adaptive cluster sampling provides estimates of the

population mean with smaller variance (which translates to lower cost for a specific degree of precision) than simple random sampling.

Thompson and Seber (1996) use the following cost model for adaptive cluster sampling with $n_1$ units in the initial sample and $L$ units in the final sample. The cost components are:

$C_T$ = total cost
$C_0$ = fixed cost, independent of sample sizes (initial or final)
$C_1$ = marginal cost per unit in the initial sample
$C_2$ = marginal cost per unit added after the initial sample

The total cost for a fixed set of initial and final sample units is given by $C_T = c_0 + c_1 n_1 + c_2 (L - n_1)$. However, since $L$ is random then the total cost $C_T$ is also random. The expected total cost is given by:

$$E(C_T) = C_0 + C_1 n_1 + C_2 (E(<) - n_1) = C_0 + (C_1 - C_2) n_1 + C_2 E(<) \qquad (9A\text{-}6)$$

In some situations, it costs more to observe a unit that contains information about the character of interest than one that does not. For example, suppose the character of interest is whether a contaminant threshold value is exceeded. If an initial unit is found to have some level of contamination using a quick field measurement, an additional measurement is made to determine the actual level. It would take longer and cost more to observe a unit with a nonzero level than a unit with no contamination. Chapter 5 of Thompson and Seber (1996) states that if comparisons of sampling strategies are to be made on the basis of cost, then "the relative advantage of the adaptive to the nonadaptive strategy would tend to be greater than in comparisons based solely on sample size."

**OPTIMUM SAMPLE SIZE FOR ESTIMATION OR HYPOTHESIS TESTING**

It is difficult, to evaluate the performance of a statistical test of hypothesis when using adaptive cluster sampling. For this reason, it is also difficult or impossible to determine optimal numbers of samples (Step 7 of the Data Quality Objectives Process). In estimating the mean, the final sample size is a random quantity and so cannot be determined in advance. However, in the guidelines proposed in Chapter 5 of Thompson and Seber (1996) for increasing the efficiency of adaptive cluster sampling (see previous section for a summary of factors), a general idea is provided of how the sample size should be distributed between the edge units and the units satisfying the condition, and how much larger the final sample size should be relative to the initial sample size. Furthermore, if the extent and abundance of the population are underestimated, one could end up with more units than time or cost would allow. If additional sampling is curtailed because of this underestimation, biases can occur in the estimation of the mean. Thompson and Seber (1996) address this issue and offer suggestions on limiting total sampling effort. As an alternative to adaptive cluster sampling, the technique of kriging may

be considered when estimating the overall mean and variance [for example, see Isaaks and Srivastava (1989)].

# CHAPTER 10

# COMPOSITE SAMPLING

## 10.1 OVERVIEW

Composite sampling involves physically combining and homogenizing environmental samples or subsamples to form a new sample (i.e., a composite sample). The chemical or biological analyses of interest are then performed on (aliquots of) the composite sample. Because the compositing physically averages the individual samples, averaging the analytical results of a few composites can produce an estimated mean that is as precise as one based on many more individual sample results. Since fewer analyses are needed, composite sampling can substantially reduce study costs when analysis costs are high relative to the costs associated with the collection, handling, and compositing of the samples. Depending on the particular situation, the particular units that comprise a composite sample may or may not need to have resulted from a prescribed sampling design (for example, random or grid sampling). Unlike the other sampling approaches described in this guidance, composite sampling does not refer to a statistically based strategy for selecting samples. Although composite sampling may assume that the samples have been collected according to such a design, it also assumes that the compositing be carried out according to some protocol (i.e., a composite sampling protocol). The composite sampling protocol identifies how the compositing is to be carried out (for example, which samples are used to form each composite) and how many composite samples are to be formed. For example, Figure 10-1 shows a situation in which $n$=9 individual samples are selected and are used to form $m$=3 composites of $k$=3 samples each; hence only $m$=3 analyses, one per composite, are needed.

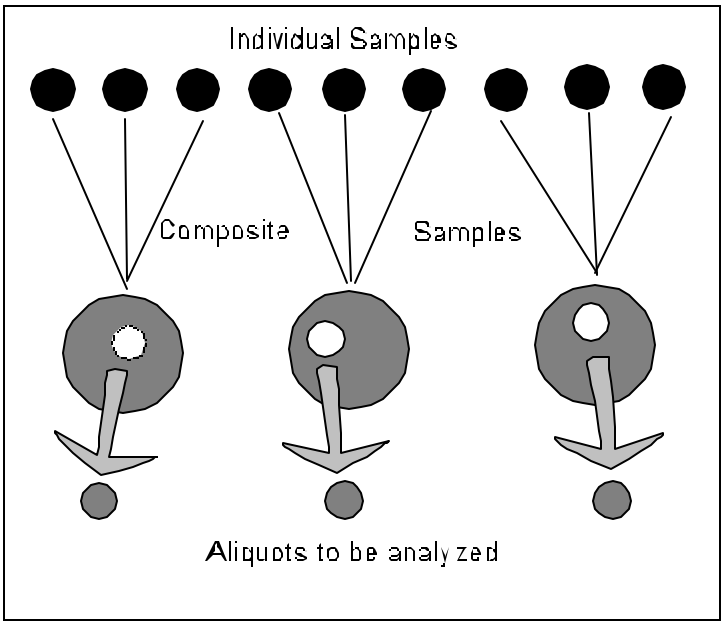


**Figure 10-1. Equal Volume, Equal Allocation Compositing**

In the designs considered elsewhere in this document, there is a one-to-one correspondence between the number of samples taken and the number of laboratory analyses performed. In cases where composite sampling can be used, however, there are n samples taken, but only $m<n$ analyses

are generally performed, one for each composite. There are several fundamental ideas associated with composite sampling:

- A primary goal is to reduce cost by having fewer analyses.
- The sample acquisition and handling process can be separated from the measurement process.
- Compositing results in a physical averaging of the samples making up the composite so that:
  - If the concentrations of a contaminant could be measured accurately in the individual samples as well as in their composite, and if the compositing process is carried out properly, then we would expect the measured level for the composite sample to be equal to the average of the measurements made on the individual samples (assuming no measurement errors).
  - Variability among similarly formed composite samples is less than the variability of the individual samples.
  - Composite sampling is naturally compatible with a study goal of estimating a population mean, while other goals may not be compatible with composite sampling, since some information is lost.

Although composite sampling has historically been used mainly for estimating a mean, it can also be used in some cases to estimate the proportion of a population that has a particular trait.

Another use for composite sampling is in the identification of a rare trait—that is, in classifying samples as having or not having a trait. In this case, only aliquots of the individual samples are composited so that some of the individual samples can be retested, based on the analytical results found for the composite. If any unit comprising a composite has the trait, then the composite will as well. Since composite sampling and retesting is aimed at classifying each unit, as opposed to making a statistical inference about the population represented by the sample units, composite sampling in this context will typically be used when there is a finite number of units that need to be classified. Suppose, for example, that water samples have been collected from all the drinking water wells within some region, and it is suspected that a small proportion of these $n$ wells are contaminated (in some specific way). Rather than testing each individual sample ($n$ tests), it may be adequate to group the samples into sets of eight. Within each group, aliquots of equal volume are taken and a test is made on the composite. If the test shows contamination, then aliquots from the individual samples are tested to identify the specific ones that are contaminated; if a small proportion of wells are contaminated, then many of the composites will yield negative results. Thus, such a strategy, or even more sophisticated

ones, can substantially reduce the amount of testing. This use of composite sampling and retesting is sometimes called group testing or screening.

In some circumstances, composite sampling can be coupled with retesting to identify the particular unit that has the highest level of a contaminant or the set of units having the highest levels (for example, those in the upper 1% of the distribution). Like the prior situation, this would often be applied when there is a finite number of units of interest. These strategies assume that measurement errors are negligible. The relative magnitudes of the composite sample measurements are used to decide on which composite(s) could potentially contain the unit with the highest level; the individual samples within those composites are then retested to determine which unit has the maximum concentration. Table 10-1 identifies the four main situations for which or composite sampling and retesting are meaningfully applied and shows where these are discussed.

**Table 10-1. When to Use Composite Sampling — Four Fundamental Cases**

| Type | Objective | Section |
|---|---|---|
| Objectives that rely on composite sampling | Estimating a population (or stratum) mean for a continuous measure (for example, analyte concentration)* | 10.2 |
| | Estimating proportion of population exhibiting some trait | 10.3 |
| Objectives that rely on composite sampling and retesting protocols | Classifying sampling units as having or not having some trait such as a being in a hot spot or from a contaminated well | 10.4 |
| | Identifying the sampling unit with highest value of some continuous measure (for example, concentration), or identifying sampling units in the upper percentiles | 10.5 |

* In general, information on variability and spatial (or temporal) patterns is lost when compositing is used for this objective; however, in some cases, some information on patterns can be acquired.

In the first two cases (Sections 10.2 and 10.3), there is interest in *making an estimate for a prescribed target population*—in the first case estimating the mean of a continuous measure (e.g, the mean concentration of contaminant) and in the second case estimating the proportion of the population with a characteristic. In these two cases, carrying out the Composite sampling means combining a sampling design with a compositing protocol. The sampling design describes the method for selecting

units from the target population and indicates the number of units to be selected and which ones are to be selected. The compositing protocol describes the scheme for forming and processing (mixing and homogenizing) composites. It indicates whether entire samples or aliquots are to be combined, the number of groups of units to be formed ($m$), the number of units per group ($k$), which units form each group, and the amount of material from each unit to be used in forming the composite sample.

The last two cases (Sections 10.4 and 10.5) involve *decision making at the unit level* rather than at the target population level. As a consequence, these approaches involve composite sampling and retesting protocols that not only define how composites are to be formed but also define when and how subsequent testing is to be done to ultimately identify particular units. The retesting strategies for these cases are conditional on the results obtained for the composites. In order to retest individual samples, the identity and integrity of the individual samples must be maintained; this implies that aliquots from the individual samples, rather than the whole samples, must be combined in forming composites. Additional aliquots from the individual samples are then retested either singly or in other composites.

Before considering composite sampling for one of the purposes indicated above, careful consideration needs to be given to its advantages and disadvantages. These are discussed in more detail in the subsequent sections; however, Table 10-2 provides some general guidance as to when Composite sampling may be useful and practical. In addition to its other potential merits, it should be noted that Composite sampling may sometimes be needed in order to have an adequate mass for analysis (for example, for dust samples or tissue samples). Finally, it should be noted that a single investigation may have several objectives —for example, estimating a population mean and its precision, as well as identifying, through retesting, the units with the highest levels. Innovative ways of applying Composite sampling need to be considered in these circumstances as a means of achieving major cost savings.

Useful references for understanding compositing and its various uses are Garner et al. (1988) and Patil et al. (1996). For estimating a mean, additional material is provided in Chapter 7 of Gilbert (1987).

## 10.2 COMPOSITE SAMPLING FOR ESTIMATING A MEAN

### 10.2.1 Overview

This section discusses Composite sampling when the objective of the study is to estimate a population mean (for example, an average site or process concentration of a contaminant). The focus is on the following situation:

**Table 10-2. Criteria for Judging Benefits of Composite Sampling**

| Criterion | Composite sampling is likely to be beneficial if... |
|---|---|
| Analytical costs | Analytical costs are high relative to sample acquisition/handling costs. |
| Analytical variability | Analytical variability is small relative to inherent variability of the site or process. |
| Objective is to estimate population mean | Information on individual samples is not important.<br>Information on associations is not important (for example, correlations between concentration levels of two contaminants). |
| Objective is to estimate proportion of population with a trait | Composite has trait if individual sample does.<br>Likelihood of misclassification is small.<br>Trait is rare. |
| Objective is to classify samples as having/not having a trait | Composite has trait if individual sample does.<br>Likelihood of misclassification is small.<br>Retesting of aliquots from individual samples is possible.<br>The trait is rare. |
| Objective is to identify the sample(s) with the highest value | Measurement error is negligible.<br>Retesting of aliquots from individual samples is possible. |
| Range of analytical concentrations | Concentrations of relevance are much larger than detection limits. |
| Physical barriers | - Compositing does not affect sample integrity (expect no chemical reactions/interferences or analyte losses from volatility) or result in safety hazards.<br>- Individual samples can be adequately homogenized. |

1. The individual samples comprising the composite are of equal size (in volume or mass) and shape.

2. The number of samples comprising each composite is the same.

3. A single subsample or aliquot is selected for analysis.

4. A single analysis is performed on the subsample.

5. A large number of composite samples could potentially be formed, but the number to actually be formed is small relative to the number of potential composites.

These are the most common conditions. Condition (1) above is necessary if the composite is to be considered equivalent to a simple averaging of the individual samples. Condition (2) is desirable so that all properties of all the composites will be the same. Conditions (3) and (4) would generally be used when compositing is contemplated, since the variation associated with aliquots and measurements would usually be smaller than that inherent to the population. Chapter 7 of Gilbert (1987) discusses the modifications to the computations of estimates and statistical analysis that are needed if the units are not of equal size or if conditions (3) or (4) or (5) are relaxed.

### 10.2.2 Application

Composite sampling for estimating a mean will generally be an appropriate strategy when all of the following conditions hold:

1. The anticipated levels for most composites will exceed detection limits so that difficulties of mean estimation in the presence of non-detects are avoided.

2. Compositing will not affect the sample integrity.

3. There are no other goals that conflict with the notion of compositing. In particular, information regarding levels for individual samples, their spatial or temporal locations, and their population variability is not considered important; information on associations (for example, correlations of concentrations of two constituents) is also not considered important.

4. Analytical costs are high relative to costs associated with sampling, sample acquisition, sample handling, and compositing. Otherwise, composite sampling will not be cost-effective.

5. There are no practical difficulties that impede the selection of multiple samples of units, where each sample is selected according to a given statistical design (for example, a simple random sample or a ranked set sample).

6. There are no practical difficulties in forming appropriate composites (for example, individual samples can be adequately homogenized).

Two comments concerning condition (1) need to be made. First, since information on variance is lost when samples are composited, repeated application of the composite sampling protocol may be needed to get information on variability that is needed to conduct a hypothesis test or to form a confidence interval. Under these circumstances, there is an implied secondary objective of the study—namely, estimation of the standard error of the estimated mean. Second, some special composite sampling protocols can be employed along with innovative data analysis techniques so that some information on spatial patterns can be resolved. Lancaster and Keller-McNulty (1998), for instance, give an example in which a site is partitioned into rows and columns to form squares and for which composite samples were formed for all squares within each row and for all squares within each column. By using statistical modeling approaches (analysis of variance type concepts) to predict areas likely to have levels of a contaminant (a secondary objective in addition to mean estimation), some information on the spatial patterns was obtained, while estimation of the mean and its standard error, the primary objectives were also accomplished by appropriately averaging the results from the composite samples.

### 10.2.3 Benefits

The main benefit of using composite sampling to estimate a mean is that one can achieve approximately the same precision of an estimated mean at less cost or, one can get more coverage (better representation) of the population at the same cost (see also Box 10-1).

A second benefit of composite sampling is that data analysis is usually easy. For example, if composite sampling is coupled with simple random sampling or grid sampling (where each composite represents the entire population), and samples of equal volumes are used to form the composites, then the simple average of the m composite samples provides an estimate of the target population mean. Section 7.3 of Gilbert (1987) and Elder, Thompson, and Myers (1980) describe procedures when equal volumes or allocations are not used.

### 10.2.4 Limitations

The main limitations are those implied by the conditions [other than (2)] delineated in Section 10.2.2. As noted, composite sampling yields a reduced amount of information on variability. If the composite sampling protocol allows each composite to represent the entire target population, then the measured concentration of the contaminant provides an estimate of the population mean concentration, assuming the measurement process is unbiased and that the compositing has been carried out properly. However, this process is repeated several times to estimate the precision of the estimated mean. Composite sampling also loses information on individual samples, spatial (or temporal) patterns, and other associations. For instance, if the compositing is over time, then information on when the contaminant had high concentrations is lost; similarly for space. Temporal and spatial correlations often

## Box 10-1. Example of Benefits of Composite Sampling

Assume zero cost for compositing and no additional variance is introduced due to composting. Suppose k = 5, m=20 (i.e., 20 composites of 5 units each). Assume the ratio of analysis to sampling costs is as shown below; then the costs achieved by compositing relative to the cost without compositing (i.e., relative to a simple random sampling of 100 units) is as shown (fixed set-up costs are ignored):

| Ratio of per-unit analysis cost to per-unit sampling/handling cost: | 2 | 3 | 4 | 5 | 10 | 20 |
|---|---|---|---|---|---|---|
| Relative cost of study: | 0.47 | 0.40 | 0.36 | 0.33 | 0.27 | 0.24 |

Note that the relative cost approaches 1/k as the ratio gets larger. Assume the ratio of measurement error standard deviation ($F_M$) to inherent variation standard deviation ($F_I$) is as shown below; the precision of the estimated mean relative to the mean determined without compositing is as shown:

| Ratio $F_M/F_I$: | 0.01 | 0.05 | 0.1 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|---|---|---|
| Std. error of mean*: | 1.000 | 1.005 | 1.020 | 1.074 | 1.153 | 1.246 |

* relative to an simple random sampling of 100 units, without compositing.

Suppose k = 10, m=10 (i.e., 10 composites of 10 units each). Assume the ratio of analysis to sampling costs is as shown below; then the costs achieved by compositing relative to the cost without compositing (i.e., relative to a simple random sampling of 100 units) is as shown (fixed set-up costs are ignored):

| Ratio of per-unit analysis cost to per-unit sampling/handling cost: | 2 | 3 | 4 | 5 | 10 | 20 |
|---|---|---|---|---|---|---|
| Relative cost of study: | 0.40 | 0.33 | 0.28 | 0.25 | 0.18 | 0.14 |

Note that the relative cost approaches 1/k as the ratio gets larger. Assume the ratio of measurement error standard deviation ($F_M$) to inherent variation standard deviation ($F_I$) is as shown below; the precision of the estimated mean relative to the mean determined without compositing is as shown:

| Ratio $F_M/F_I$: | 0.01 | 0.05 | 0.1 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|---|---|---|
| Std. error of mean* | 1.000 | 1.011 | 1.044 | 1.160 | 1.320 | 1.497 |

* relative to an Simple random sampling of 100 units, without compositing.

exist between concentrations of two contaminants, but this information is also lost when composites are used.

Composite sampling may be difficult to accomplish without introducing large additional errors (in weighing or homogenizing). This is especially true for a solid medium such as soil or dust if they are heterogeneous. Homogeneity in this context refers to characteristics of the environmental medium that affect one's ability to form a "good" composite. In general, composite sampling works better for more homogeneous media. As Pitard (1993) notes, however, homogeneity is a relative concept: sand particles seen from a distance may seem homogeneous but the individual grains will reveal great heterogeneity in shape, size, color, density, and so forth. In compositing, the $k$ individual samples are mixed to the point that one can expect that $1/k$ of any aliquot came from each individual sample. Achieving this can be especially difficult if the individual samples are dissimilar in properties that affect homogenization. For example, soil composites comprised of some samples that are mostly clay and some that are mostly sand will tend to be poor candidates for composite sampling. On the other hand, liquids tend to be more homogeneous than solids so that liquids are typically good candidates for composite sampling, but even liquids must be thoroughly mixed prior to compositing, if aliquots are to be composited, and prior to taking aliquots from the composite.

### 10.2.5 Implementation

Box 10-2 provides directions for implementing composite sampling to estimate a mean in the simplest situation—namely, when equal-volumes and equal-allocations of samples or aliquots are used to form composites and where each composite represents the entire target population (by virtue of random sampling, for instance).

From Box 10-2 note that steps 1, 2, and 3 involve considerations for arriving at appropriate values for $k$ and $m$; details are provided in Appendix 10-A. The basic mechanism for determining $k$ and $m$ is to develop a cost model that expresses the total cost in terms of its per-unit and per-analysis costs and to develop a variance model that expresses the variance of an estimated mean in terms of variance components:

$$\text{Total Cost} = \text{fixed set-up costs} + n(\text{cost of sample collection/handling}) + m(\text{cost of analysis})$$

$$\begin{aligned}\text{Variance (mean)} &= \text{Variance [sum of } m \text{ observed concentrations}/m] \\ &= [\text{inherent concentration variability} + k\ (\text{error variability})]/n\end{aligned}$$

where $n = mk$.

**Box 10-2. Directions for Selecting Equal Allocation, Equal Volume Composite Samples for Estimating a Mean**

STEP 1: If compositing costs can be considered negligible relative to other costs and additional variability introduced by the compositing process can be ignored, use Table 10-3 to determine an initial $k$ value. Determine an appropriate volume for each sample, and determine if the indicated $k$ can be used, based on the physical nature of the samples, the anticipated levels of the contaminant(s) of interest (relative to detection limits), and the capability to combine and homogenize them adequately. If the indicated $k$ is too large based on these practical considerations, use the maximum practical $k$.

STEP 2: Choose $m$ so that selection of $n = mk$ samples and analysis of m samples will be within the budget: Cost equals the fixed set-up costs plus the cost of selecting/handling n samples plus the cost of forming m composites plus the cost of performing m analyses.

STEP 3: Check that $m$ is large enough to produce a sufficiently precise estimate of the mean: Variance of estimated mean equals the variance of mean without compositing plus measurement error variance multiplied by $(k-1)/mk$ (assuming compositing adds no error). If the precision needs to be improved, or if confidence intervals or tests are to be performed for the mean (which may imply the need for a larger $m$), then consider cost-precision tradeoffs in which m is increased and $k$ is reduced.

STEP 4: Select $k$ samples according to the prescribed sampling design and compositing protocol. (See Chapter 4 for procedures for selecting a random sample.)

STEP 5: Form one composite sample as follows: physically mix and homogenize each sample (this may be unnecessary if each sample is of the same volume and each sample is to be included in the composite in its entirety). Obtain equal-volume aliquots of material from each of these k samples and physically mix the selected material and homogenize it thoroughly.

STEP 6: Repeat steps 4 and 5 m times to form m composite samples.

STEP 7: Obtain measurements on the environmental parameters of interest for each composite sample or, if necessary, on equal-volume aliquots therefrom.

**Table 10-3. Optimal *k* Values for Estimating a Population Mean***

| Cost Ratio $C_M/C_S$ | Ratio of Std. Dev. of Measurement Errors to Std. Dev. of Inherent Levels | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0.10** | **0.20** | **0.30** | **0.40** | **0.50** | **0.60** | **0.70** | **0.80** | **0.90** | **1.00** |
| **2** | 14 | 7 | 5 | 4 | 3 | 2 | 2 | 2 | 2 | 1 |
| **3** | 17 | 9 | 6 | 4 | 3 | 3 | 2 | 2 | 2 | 2 |
| **4** | 20 | 10 | 7 | 5 | 4 | 3 | 3 | 3 | 2 | 2 |
| **5** | 22 | 11 | 7 | 6 | 4 | 4 | 3 | 3 | 2 | 2 |
| **8** | 28 | 14 | 9 | 7 | 6 | 5 | 4 | 4 | 3 | 3 |
| **10** | 32 | 16 | 11 | 8 | 6 | 5 | 5 | 4 | 4 | 3 |
| **15** | 39 | 19 | 13 | 10 | 8 | 6 | 6 | 5 | 4 | 4 |
| **20** | 45 | 22 | 15 | 11 | 9 | 7 | 6 | 6 | 5 | 4 |
| **50** | 71 | 35 | 24 | 18 | 14 | 12 | 10 | 9 | 8 | 7 |

* The entries assume that compositing costs are negligible, that any increase in variability due to compositing is negligible, and that the samples to be composited are selected at random from the entire sample. $C_M$ and $C_S$ are the per-unit analysis cost and the per-unit sample acquisition/handling cost, respectively.

The "inherent concentration variability" component refers to the natural variability in the true concentrations that occurs among units in the target population, whereas the "error variability" component refers to variability resulting from random errors made in the sample collection and measurement processes. The variance formula above assumes that the individual samples comprising each composite come from independent simple random samples of the entire population. To arrive at corresponding formulae for a simple random sampling without compositing, one can substitute $m=n$ and $k=1$ into the above. The above formulae are simplified since they omit the cost of forming composites, which would typically depend on $k$, and the additional variance that might result from compositing.

Under these simplifying assumptions, an optimal $k$ can be determined as a function of the relative cost components and variance components (see Appendix 10-A). A table of optimal $k$ values for this situation is furnished in Table 10-3. Prior knowledge about the population variability and the anticipated error variability is needed to use this table. For instance, suppose a prior study investigating soil contamination resulted in an observed distribution of concentrations of the contaminant of interest that had a coefficient of variation (CV) of 90%; based on geological considerations and the fact that the study under consideration plans to use the same measurement methods, a similar CV is anticipated. Analysis of the concentrations from a series of duplicate samples (i.e., quality control data from the prior study) revealed a measurement-error relative standard deviation of 18%.

The measurement error variance thus accounts for $100(18/90)^2$ % of the total variance, or 4%; hence the inherent variation is anticipated to be 96% of the total variance. The ratio of the standard deviation of measurement errors to the standard deviation of inherent levels can therefore be determined as the square root of (4/96), or approximately 0.20; this identifies the column in Table 10-3 to be used. Since the per-unit measurement cost is expected to be about three times the cost of sample acquisition, the table shows that using k=9 will be optimal (from a statistical standpoint) for the planned study. As noted above, this assumes both that the compositing of nine samples is practical and that components of cost and variance associated with the compositing process can be ignored.

Box 10-3 provides a hypothetical example that demonstrates the benefits of compositing in terms of costs and precision by comparing the results from a particular compositing scheme to those that would have been obtained from a grid sample.

### 10.2.6 Relationship to Other Sampling Designs

In addition to coupling composite sampling to a simple random sample or a grid sample, several other alternative composite sampling protocols may be worth considering. These include the following:

- If *k* strata of equal size (volume or mass) can be used, then one sample per stratum can be picked and those *k* samples composited. This can avoid having a random sample that is clustered in a small subregion of the target area. If this process is repeated *m* times, then the simple average of the composite sampling results will estimate the target population mean, and the variability of the composite sampling results can be used to estimate the precision of the mean. The precision of the mean via this approach may be poorer than that from an simple random sampling, but one is assured that the sample achieves adequate coverage of the target population. If the stratifying criteria are correlated with the variable of interest (the rationale for stratifying in the first place), then the precision of the mean is increased.

- If compositing within strata is used rather than compositing throughout the site or process, then the precision of the estimated mean may be improved over that achieved by coupling composite sampling with a simple random sample or grid sample from the overall population. This improvement should be realized if the contaminant concentrations are more homogeneous within strata than among strata. A major advantage of this approach is that some spatial or temporal information is attained; also, if retesting is feasible, then information on selected specific units can be efficiently acquired. But a drawback is the difficulty in getting a good estimate of the precision of the mean, especially if the variability within the different strata is different. This drawback occurs because there must be enough composites per stratum to obtain separate variance estimates for each stratum; rather than having a single overall

## Box 10-3. Example: Compositing for Estimating a Site Mean

The goal was to estimate the average surface soil concentration of contaminant X for a 30 meter x 40 meter site. The site was partitioned into 12 square (10 meter x 10 meter) subareas (cells). Assume that three randomly sited grid samples of 12 points each were placed over the site (*i.e.,* one point per grid cell) and that the observed X concentrations for *individual samples* were as follows:

| | | | | |
|---|---|---|---|---|
| Grid 1: | 853 | 986 | 1090 | 2344 |
| | 885 | 1082 | 767 | 1592 |
| | 528 | 763 | 993 | 806 |
| Grid 2: | 983 | 869 | 1740 | 258 |
| | 799 | 756 | 643 | 1747 |
| | 794 | 751 | 985 | 1106 |
| Grid 3: | 1161 | 886 | 1256 | 2276 |
| | 838 | 791 | 1267 | 2034 |
| | 714 | 907 | 913 | 1118 |

Data Analysis Results and Costs for Each Grid Sample (and Combined)—Based on Individual Sample Analyses

| Grid Number | 1 | 2 | 3 | Combined |
|---|---|---|---|---|
| Observed Mean Conc. | 1057 | 1147 | 1180 | 1128 |
| Std. Error | 139 | 168 | 142 | 85 |
| 95% Confidence Interval | (752, 1362) | (777, 1517) | (867, 1493) | (960, 1296) |
| Costs: | | | | |
| - for collecting/handling | 12x$25=$300 | 12x$25=$300 | 12x$25=$300 | 36x$25=$900 |
| - for analysis | 12x$500=$6000 | 12x4500=$6000 | 12x$500=$6000 | 36x$500=$18000 |
| - total | $6300 | $6300 | $6300 | $18900 |

Data Analysis Results and Costs for Each Grid Sample (and Combined)—Based on Composite Sample Analyses

| Grid Number | 1 | 2 | 3 | Combined |
|---|---|---|---|---|
| Observed Mean Conc.* | 1070 | 976 | 1125 | 1057 |
| Std. Error | | | | |
| 95% Confidence Interval | | | | (865, 1249) |
| Costs: | | | | |
| - for collecting/handling | 12x$25=$300 | 12x$25=$300 | 12x$25=$300 | 36x$25=$900 |
| - for compositing | 1x$20=$20 | 1x$20=$20 | 1x$20=$20 | 3x$20=$60 |
| - for analysis | 1x$500=$500 | 1x$500=$500 | 1x$500=$500 | 3x$500=$1500 |
| - total | $820 | $820 | $820 | $2460 |

* Assumes the same analytical error variability as for the individual samples; also assumes no additional error occurs as a result of the compositing process.

The above data were generated for a site in which the true site mean was 1113. The true means of the three grids were 1063, 1149, and 1174, respectively, for a true grid average of 1129.

estimate of variance, one in this case forms a weighted combination of the separate estimates to arrive at the estimated precision for the mean.

The overall variance estimator takes the form of the variance of the overall mean obtained with a stratified sampling design (see Appendix 6-A of Chapter 6).

To produce a meaningful estimate of a population mean, composite sampling is suitably linked with some type of sampling design. Some examples follow:

**Simple random sample.** In this case, *m* simple random samples of *k* units each are selected from throughout the site/process, and the units in each sample are used to form composites. In this case, the estimate of the population mean equals the simple average of the *m* composite-sample measurements, if equal-volume samples or aliquots are used. The standard error of the mean (with the same assumption) is calculated based on the variability among the *m* composites. If confidence intervals or hypothesis tests are to be performed (which depend on having a good estimate of the standard error, as well as the mean), then there must be adequate *m* for estimating the variance among the composites. To achieve this, some tradeoff of *m* versus *k* values may be needed.

**Grid sample.** Composite sampling in this case would involve forming *m* grids having different starting values; each grid would have *k* points, and samples from those points would be composited. The estimation of the mean and its variance would be like that for a simple random sample. See the example in Box 10-3.

**Stratified random or stratified grid sample.** If composites are formed from samples within strata, then the target population mean is estimated as a weighted average of the stratum means. The variability of this estimated mean therefore depends on the within-stratum variation. Consequently, we would expect better precision than for a comparably sized random sample, assuming that the concentrations have less variation within strata than among strata. As noted in Section 10.2.5, however, it may be more difficult to get an estimate of the variance (or the standard error) of the mean and the cost may be higher since more composites might be needed to get such a precise estimate. The main advantage of this approach is that it retains some information on spatial or temporal variation.

**Two-stage sampling.** This is a fairly common situation. In this case, units fall naturally into groups (called "batches") or can be grouped into batches, and *h* out of *H* batches are chosen at the first stage of sampling. Then *n* samples are selected within each batch and m composites are formed from these samples. If $h=H$, then batches are equivalent to strata and this case is like the prior one. But if $h<H$, then this is considered two-stage sampling. For instance, in monitoring a wastewater stream, batches could be days, and samples selected at random times within the selected days could be used to form composites representing days. Section 7.3 of Gilbert (1987) also treats the case where multiple subsamples from each composite are extracted and multiple analyses per subsample are performed.

**Ranked set sampling.** Composite sampling can be used in conjunction with ranked set sampling. Usually this involves forming composites from units having the same ranks. In this case, the advantages and disadvantages are similar to those for stratified sampling (see above). Compositing across ranks might also be considered. For example, suppose that nine samples would be randomly selected and grouped into three groups of three each. The three samples in each group would be ranked by inspection (assumed to be correlated with the parameter of interest). The sample with rank 1 in group 1, the sample in group 2 with rank 2, and the sample in group 3 with rank 3 would be composited and analyzed. The initial group of nine samples yields only one composite sample of size 3. In terms of the precision of the estimated mean, such an approach should perform better than a simple random sample of size three (because this is characteristic of ranked set sampling) or a composite therefrom. But it should perform worse than a simple random sample of size nine or a composite of all nine samples. Hence, the compositing across ranks strategy would appear to be a good one only if there were physical or practical reasons why nine samples could not be composited but three could.

### 10.2.7 Examples

The example in Box 10-3 illustrates use of composite sampling with a grid sample. Similar approaches would apply if the cells in that example were regarded as strata, since the strata are of equal size. An alternative compositing scheme would involve compositing the samples within each of the 12 strata and averaging the results to produce an estimate of the target population mean. This strategy would be somewhat more expensive and would provide some information on spatial patterns but would not permit one to get a measure of the precision of the overall mean (unless the whole process were replicated). Depending on secondary objectives that may be of interest, other composite sampling strategies might be used that would form composites within rows and within columns.

As noted earlier, compositing sometimes provides samples with mass sufficient to permit use of chemical techniques that allow a broader range of chemicals to be detected. Such protocols may be especially useful for analyzing dust samples (for example, from various areas within a home) or certain tissue samples (for example, from several fish).

## 10.3 COMPOSITE SAMPLING FOR ESTIMATING A POPULATION PROPORTION

### 10.3.1 Overview

Under certain circumstances, composite sampling offers an efficient way to estimate the proportion of a population that has a particular trait. The goal in this case is simply to estimate the proportion with the trait; there is no interest in which units have the trait. This might be useful in the early stages of investigation when one wants to determine if further testing is warranted. For example, if fewer than 5% of the drums in a field are contaminated, then it might be worthwhile to identify them (at a later stage) and selectively remove the contaminated ones, but if more than 15% of the drums are

contaminated, it may be more cost-effective to simply remove all of them. What is needed as an initial step is an efficient way to estimate the proportion of the drums that are contaminated. In such a situation, *m* random samples of *k* drums could be selected and used to form *m* composites, and analysis of the composites would reveal whether or not they were contaminated. The analysis of each composite in this case simply yields either a positive (i.e., has trait) or a negative (i.e., does not have trait) result. A negative result for a composite implies that all of its component samples were negative (i.e., uncontaminated), while a positive result implies that one or more of the component samples was positive. When composites are formed from random samples, there is a known relationship between the proportion of composite samples expected to be positive and the proportion of individual samples expected to be positive; this relationship allows the proportion of individual samples with the trait to be estimated based on the observed proportion of positive composite samples.

### 10.3.2 Application

Use of composite sampling for estimating a population proportion, while not a particularly common practice, can be very cost-effective if the conditions for its use are appropriate. All of the following conditions should hold if composite sampling for this purpose is to be used:

- Information on individual samples and/or on their spatial or temporal locations is not considered important.

- Analytical costs are high relative to costs associated with sampling, sample acquisition, sample handling, and compositing. Otherwise, composite sampling is not cost-effective.

- The proportion of the population having the trait to be estimated is small. Otherwise, composite sampling will not be cost-effective.

- There are no practical difficulties that impede the selection of random samples of units.

- There are no practical difficulties in forming appropriate composites (for example, individual samples can be adequately homogenized).

- Compositing will not affect the sample integrity.

- The likelihood of misclassifying a composite sample is negligible. If the trait of interest is based on whether an analyte is or is not present, then this condition implies that the individual samples either should have "high" levels or (essentially) zero levels. That is, composite sampling in this context will work well when there is a clear distinction between those units with and without the trait. For instance, if a spill has occurred, then concentrations can be expected to be many times larger than the detection limits for the

affected units and essentially zero for all unaffected units. [Garner et al.(1989) describe ways of dealing with cases where misclassification errors may be non-negligible.]

### 10.3.3 Benefits

The benefit of using compositing for estimating a population proportion is the cost savings that can be achieved. As an example, suppose that 40 composite samples each comprised of 20 randomly selected units can be used. If the cost of an analysis is $100 and the cost of collecting and compositing samples is $5 each, then the total cost (excluding fixed, set-up costs) would be $8000 (40x$100 + 40x20x$5). If the (unknown) proportion of positive units in the population is 0.05, then the above design would be expected to produce a 95% confidence interval with width of ±0.021. If one chose not to use compositing, over 300 individual samples and analyses would be needed to attain about the same precision, but the cost would be substantially higher—more than $31,500 (300x$105).

### 10.3.4 Limitations

The main limitations are those implied by Conditions (4) through (7) in Section 10.3.2. Of particular concern is condition (7) when analyte presence/absence is the trait of interest, since misclassification probabilities can become non-negligible due to dilution effects. Another potential problem is that the optimal number of units to be combined into each composite, based on statistical considerations (see Section 10.3.5), may be too large to be practical [i.e., condition (5) fails to hold].

### 10.3.5 Implementation

Let $p$ denote the unknown proportion of the units in the population having the given trait (i.e., $p$ is the probability of observing a positive result for a randomly selected unit in the population). To use composite sampling to estimate $p$, one chooses $m$ random samples of $k$ units each from the population and then forms $m$ composites from the $k$ units in each sample. Each composite is then tested for the trait. If $x$ (which depends on $k$) is the number of positive test results that occur from among the $m$ composites, then $x/m$ can be used to estimate $p^*$, where $p^*$ denotes the probability that composites of size $k$ will test positive. Because the number of positive results, $x$, has a known statistical distribution (a binomial distribution with parameters m and $p^*$) when random samples are used, the relationship between $p^*$ and $p$ is $p=1 - (1-p^*)^{1/k}$. Thus by substituting the observed $x/m$ fraction for $p^*$ in this equation, an estimate for $p$ can be obtained. This estimate is satisfactory if the misclassification rates are small. Modified estimates are needed otherwise (Garner et al., 1989).

The specific directions for implementation are given in Box 10-4. As indicated there, to arrive at the appropriate combination of $m$ and $k$, one needs to have some idea of the maximal p value to be expected; also, one needs to specify the precision with which p is to be estimated. Table 10-4 provides guidance in choosing $m$ and $k$ to meet a specified precision level. Note that increasing m, the

**Box 10-4. Directions for Composite Sampling for Estimating the Proportion of a Population with a Given Trait**

STEP 1: Based on prior knowledge, determine an upper bound on *p*, the proportion of the population having the trait of interest. Because compositing will not be cost-effective when p exceeds 0.25, continue with Step 2 only if p is less than or equal to 0.25.

STEP 2: Determine an appropriate volume for each sample, and determine the maximum number of samples that can be composited, based on the physical nature of the samples and the trait to be assessed and on the capability to combine and homogenize the samples adequately. Denote this maximum number of samples as *K*.

STEP 3: Use the *p* value from Step 1 to identify a column in Table 10-4 (or an interpolated value). Within this column, consider any row in which the optimal *k* is less than *K*. Within this set of rows, select the smallest m value for which the precision is deemed acceptable. (The precision in Table 10-4 is expressed in terms of an approximate 95% confidence interval.)

STEP 4: Compute the cost of selecting $n = m \times k$ samples and analyzing *m* samples to determine if the composite has the trait of interest: Cost = fixed set-up costs + (cost of selecting/handling a sample) x *n* samples + (cost of forming a composite + cost of performing an analysis) x *m* analyses.

STEP 5: If the cost exceeds the available resources, consider cost-precision tradeoffs to arrive at useful *m* and *k* values. (Note that the restriction that k be less than or equal to *K* may lead to situations in which optimal combinations of *m* and *k* are not possible. The formulas in Appendix 10-B can be used to calculate the precision for a given combination, and the formula in Step 4 can be used to calculate the total cost.)

STEP 6: Use a simple random sampling strategy to select *k* samples at random from the target population. (See Chapter 3, for example, for procedures for selecting a random sample.)

STEP 7: Form one composite sample as follows: physically homogenize each of the *k* samples (this may be unnecessary if each sample is of the same volume and each sample is to be included in the composite in its entirety). Obtain equal-volume aliquots of material from each of these *k* samples and physically mix the selected material and homogenize it thoroughly.

STEP 8: Repeat Steps 6 and 7 *m* times to form *m* composite samples.

STEP 9: Obtain measurements on the parameters of interest for each composite sample or, if necessary, on equal-volume aliquots therefrom. Let *x* be the number of the *m* composites that have the trait and define $P=x/m$. Calculate the prevalence of the trait in the population as $\hat{p} = 1-(1-p)^{1/k}$. See Appendix 10-B for estimating precision of this estimate.

number of composites, increases the precision. The table shows the optimal $k$, from a statistical precision-of-estimation standpoint, for the different $m$ and $p$ values. Note that this optimum $k$ increases with increasing $m$. If $p$ is small, then the $k$ values indicated may be too large to be practical, either because of homogenization difficulties or because of concerns about dilution effects that would negate the assumption of minimal misclassification errors. The table also gives approximate 95% confidence interval widths that can be with the optimum $k$. Composite sampling protocols for estimating p can also be derived that take into account the relative costs of sample collection, handling, compositing, and testing. The cost model is the same as in Table 10-5 in Appendix 10-A.

### 10.3.6 Relationship to Other Sampling Designs

Simple random samples should be used in this situation; otherwise, the relationship between $p$ and $p$* may not hold.

### 10.3.7 Examples

An application of composite sampling for estimating a proportion would be seen in the early stages of a large-scale investigation that encompasses a large number of sites where each site has a defined set of sampling locations (for example, points defined by a fixed-size grid). The goal would be is to identify the subset of sites with the highest and lowest proportions (across the grid points) where further testing or remediation actions would be taken. If the individual samples can be preserved so that subsequent aliquots can be selected, then other uses of composite sampling, such as those described in Sections 11.1 and 11.2, can be used. A similar application would be a field of drums with potential contamination in some – another situation in which estimating the contamination proportion via composite sampling would be the first step in an iterative study.

**Table 10-4. Optimal *k* for Estimating *p* and Approximate Confidence Intervals for *p***

| m | Maximum Anticipated Prevalence (p) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | p=0.25 | | p=0.10 | | p=0.05 | | p=0.01 | |
| | opt. k | ~ C.I. | opt. k | ~ C.I. | opt. k | ~ C.I. | opt. k | ~ C.I. |
| 100 | 5 | ±0.06 | 14 | ±0.02 | 30 | ±0.012 | - | |
| 70 | 5 | ±0.07 | 14 | ±0.03 | 30 | ±0.015 | - | |
| 50 | 5 | ±0.08 | 12 | ±0.04 | 25 | ±0.018 | - | |
| 40 | 4 | ±0.09 | 12 | ±0.04 | 20-25 | ±0.021 | - | |
| 30 | 4 | ±0.11 | 10 | ±0.05 | 20 | ±0.025 | - | |
| 20 | 3 | ±0.14 | 8 | ±0.06 | 15 | ±0.03 | 60 | ±0.007 |
| 10 | 3 | ±0.22 | 5 | ±0.11 | 9 | ±0.06 | 35 | ±0.014 |

*Information is based on Garner et al. (1989) which contains more m & k combinations.

## APPENDIX 10-A

## COST AND VARIANCE MODELS FOR COMPOSITE SAMPLING AIMED AT ESTIMATING A MEAN

Table 10-5 provides information on the costs and variances encountered when estimating a mean either when a simple random sample is used alone or when composite sampling is coupled with random sampling. The first few rows of Table 10-5 list the typical cost components that occur. In the first (simple random sampling) case, it is assumed that $n$ samples are individually analyzed, while in the second case, it is assumed that $n = km$ samples are selected, where $m$ is the number of composites (each comprised of $k$ samples) analyzed. As indicated in the table, there are three cost components considered, in addition to the fixed set-up cost, and the total cost is determined by adding these components together, appropriately weighted by the number of samples or analyses.

The lower portion of Table 10-5 defines the pertinent variance components and indicates how they combine to yield the variance of an estimated mean for the simple random sampling and the simple random sampling with compositing cases. For the latter case, the relevant components are: (1) the inherent variability of the site or process, (2) the variability associated with collecting and handling the samples, (3) the variability associated with the compositing process, and (4) the variability associated with all aspects of the measurement process. In most circumstances, (1) and (2) cannot be separately estimated and are therefore treated as one.

From a statistical standpoint, an optimal $k$ can be determined as follows, if the cost of compositing ($C_C$) can be ignored and the additional variance due to compositing ($V_C$) can be considered negligible:

$$\text{opt.k} = \sqrt{\frac{V_I C_M}{V_M C_S}}$$

$V_I$ = the variance component associated with the inherent variability in the population plus variation due to sample collection and handling,

$V_M$ = the variance component associated with analytical measurements (on a given sample),

$C_S$ = the per-unit cost associated with collection and handling of the individual samples, and

$C_M$ = the per-unit cost associated with making an analytical measurement.

The optimum is more difficult to determine if either the cost of compositing or the additional variance due to compositing is not negligible, since these components depend on $k$. However, the formulae in the table permit one to calculate the costs and variances for any combination of $k$ and $n$.

**Table 10-5. Components of Cost and Variance for Random Samples - With and Without Composite Sampling**

| Component | Random (or Grid) Sample without Composite Sampling | Random (or Grid) Sample with Composite Sampling |
|---|---|---|
| **COSTS** | | |
| Fixed cost | $C_0$ | $C_0$ |
| Per-unit sampling cost | $C_S$ (x n samples) | $C_S$ (x n samples) |
| Per-composite compositing cost | 0 | $C_C$ (x m analyses) |
| Per-unit measurement cost | $C_M$ (x n analyses) | $C_M$ (x m analyses) |
| Total Cost | $C_0 + nC_S + nC_M$ | $C_0 + nC_S + mC_{C+} mC_M$ |
| **VARIANCES OF CONTAMINANT X** | | |
| Inherent variability* | $V_I$ | $V_I$ |
| Per-composite compositing variability | 0 | $V_C$ (= $f_C V_I$)** |
| Per-unit measurement variability | $V_M$ | $V_M$ |
| Total variance of an observed measurement | $V_I + V_M$ | $[V_I (1+f_C)]/k + V_M$ |
| Variance of mean of all measurements | $[V_I + V_M]/n$ | $\{[V_I (1+f_C)]/k + V_M\}/m =$ $[V_I (1+f_C)]/n + V_M/m =$ $[V_I (1+f_C)] + kV_M]/n$ |

where n = number of samples,
m= number of measurements (= number of composite samples), and
k = number of samples/composite.

* Includes variability associated with collecting and handling of samples.

** This component is generally small. It is considered to be proportional to the $V_I$ for the following reason: if the individual samples comprising a composite all have essentially the same X concentration (i.e., a small $V_I$) , then failure to have exactly equal volumes from each individual sample or failure to homogenize them adequately will result in relatively little error in the X level of the composite; on the other hand, these sorts of procedural flaws will lead to larger errors if the individual samples exhibit large variation (i.e., a large $V_I$). Thus, $f_c$ represents the factor (for example 0.02 for a 2% increase) which is multiplied times the inherent variability to account for variability stemming from compositing-related activities.

If an optimal $k$ value is too large for practical implementation, the maximum practical $k$ would typically be employed. Further, a smaller $k$ value might be employed if a sufficiently large $m$ is needed to estimate adequately the precision of the estimated mean. With the "optimal" $k$, one would expect to obtain a more precise estimate but may not be able to estimate the precision well. Hence in those cases where a good estimate of the precision is also needed—for example, if confidence interval estimates are to be constructed or hypothesis tests are to be performed (since these rely on both an estimate of the mean and its standard error)—one should consider using a smaller $k$ if that is the only way to achieve an adequate number of composites, $m$, for estimating the standard error.

Box 10-1 of Section 10-2 provides another example of how the information in Table 10-5 can be used. It compares the cost and precision for a composite sampling approach versus an simple random sampling, under the assumptions that $C_C$ and $V_C$ are negligible. The cost comparison is derived by dividing the total cost in the last column by the total cost in the first column. After some algebra, this relative cost ratio can be shown to be:

$$RC = \frac{C_0 + nC_S\left[1 + \frac{1}{k}\frac{C_M}{C_S}\right]}{C_0 + nC_S\left[1 + \frac{C_M}{C_S}\right]}$$

If the fixed-cost component is ignored (as in Box 10-1), then this equation shows that the relative cost, RC, reduces to a function of k and the $C_M/C_S$ ratio (i.e., only the part inside the brackets remains). The precision comparison given in Box 10-1 was derived from the information in the last row of the table; after some algebra, and the assumption that $f_C = 0$, the relative variance (RV) of composite sampling versus an simple random sampling without compositing is:

$$RV = \frac{1 + k\left(\frac{V_M}{V_I}\right)}{1 + \left(\frac{V_M}{V_I}\right)}$$

## APPENDIX 10-B

## COMPOSITE SAMPLING FOR ESTIMATING A POPULATION PROPORTION

Details for using composite sampling to estimate a proportion are provided by Garner et al. (1989). It is assumed that there is a large (essentially infinite) population of units and that n are selected at random. Suppose $m=n/k$ composites of $k$ units each are formed (by random groupings). Let the probability of a composite testing positive be denoted by $p^*$ and let the probability of an individual unit testing positive be denoted by $p$ ($p$ is the proportion of the population having the trait and is the quantity that is to be estimated). Since the probability of a composite testing negative is the same as the probability that all $k$ units will test negative, then:

$$1 - p^* = (1 - p)^k \text{ or } p^* = 1 - (1 - p)^k \qquad \text{(10B-1)}$$

Solving for $p$ yields

$$p = 1 - (1 - p^*)^{1/k} \qquad \text{(10B-2)}$$

This equation offers a simple way to estimate $p$: If $x$ of the $m$ composites are found to have the trait, then substituting $x/m$ for $p^*$ in this equation will yield an estimate of $p$. This estimate is a maximum likelihood estimate; it is not an unbiased estimate, however.

Since the number of positive composites has a binomial distribution with parameters $m$ and $p^*$, one can use tables for binomial confidence limits to derive corresponding confidence limits for $p$. [Binomial confidence limits are available in many standard statistical texts; Gilbert (1987), also provides some tables and formulae.] For example, suppose composites comprised of 10 units each were used. If 10 out of 30 of these composites tested positive, then the point estimate for $p^*$ is $x/m$=0.3333 and a 95% confidence interval for $p^*$ is [0.1729, 0.5280]. Using $x/m$ and these interval end points in equation 10B-2 yields a point estimate of $p$ equal to 0.0397 and a corresponding interval estimate of [0.0188, 0.0723].

If the mis-classification rates are not negligible, and one knows them, then one can substitute:

$$\frac{\frac{x}{m} - \alpha}{1 - \alpha - \beta}$$

for the $p^*$ appearing in equation 10B-2, where " and $ are the false rejection and false acceptance rates, respectively.

# CHAPTER 11

# COMPOSITE SAMPLING FOR IDENTIFYING A TRAIT AND EXTREME SAMPLING UNITS

## 11.1 COMPOSITE SAMPLING FOR IDENTIFYING A TRAIT

### 11.1.1 Overview

Another main use for composite sampling is to classify units as having or not having a trait. In this case, some of the units are retested; hence, the identity of the units is maintained and only aliquots (not the entire samples) of the individual sampled units composited. The particular units to be retested depend on the analytical results found for the composites. Hence the analytical results must be readily available for these composite sampling and retesting schemes to be practical. Since the goal of the composite sampling and retesting protocol in this situation is to classify each unit, as opposed to making a statistical inference about the population represented by the units, composite sampling and retesting would typically be encountered when there is a finite number of units and classification of all units is needed. This section discusses various composite sampling and retesting schemes. There are two basic cases to distinguish. A fundamental assumption *for both cases* is that *measurement error is very small* and does not interfere with the yes/no identification.

The first case involves a situation in which a truly binary trait (a yes/no measure) is the trait of interest. Using the schemes in this case assumes that if any unit comprising a composite has the trait, then the composite will as well. As a consequence, samples in a "negative composite" do not need additional testing, while samples in a "positive composite" need further testing via some scheme. In this situation, there is a clear distinction between units having and not having the trait—i.e., misclassification errors (including, for instance, impacts of dilution effects) are very small. The fundamental notion underlying this use of composite sampling is that by retesting units (either singly or in subcomposites), the identity of the positive unit(s) can be determined and if the trait is rare enough, then the composite sampling and retesting strategy will need fewer analyses than simply testing all individual units. The literature for this case is extensive, and determination of properties of various composite sampling and retesting schemes (for example, expected cost, as a function of prevalence of the trait) is relatively straightforward, at least for the simpler schemes.

The second case involves comparing a continuous, non-negative measure $x$ (for example, a concentration of contaminant) with a threshold level $C$. (For ease of exposition, $x$ will be referred to as a concentration.) In this case, a unit having $x \geq C$ is said to have the trait. To identify such units when composites are employed, one obviously has to adjust the threshold level. The worst case situation would be if $x=C$ for one unit in a composite and $x=0$ for the remaining $k$-1 units; to identify the unit with $x=C$, one would need to compare a composite's concentration to $C/k$, if $k$ samples are used to form

composites. The difficulties with this case are twofold: (1) a "positive composite" does not necessarily imply that one or more of its component units is positive (i.e., there can be composites with concentrations greater than C/k that do not contain any individual units with x $C) and (2) the assumption of negligible measurement errors is likely to be untenable in many circumstances. It is much more difficult to estimate costs for a given composite sampling and retesting scheme in this case because the expected number of analyses depends on the spatial (or temporal) distribution of the underlying measure x rather than just on the prevalence of the trait. Annotated references for composite sampling and retesting are provided by Boswell et al. (1992).

### 11.1.2 Application

It will generally be appropriate to use composite sampling and retesting for classifying units when the following conditions hold:

- There is a predefined set of units that are to be classified. This could be a sample from some population, but inferences are limited to the units actually within the set.

- Analytical costs are high relative to costs associated with sampling, sample acquisition, sample handling, and compositing. Otherwise, composite sampling and retesting will not be cost-effective.

- The proportion of the population having the trait of interest is small. Otherwise, composite sampling and retesting will not be cost-effective.

- Representative aliquots from the individual units can be obtained.

- There are no practical difficulties in forming appropriate composites from aliquots (for example, aliquots can be adequately homogenized).

- Compositing will not affect the sample integrity.

- Retesting of units is feasible. In particular, the identity of the units can be maintained, and samples can be adequately preserved throughout all the potential stages of testing and retesting.

- The likelihood of misclassifying a composite sample is minimal. If the trait of interest is based on whether an analyte is or is not present, then this condition implies that the individual samples either should have "high" levels or (essentially) zero levels—i.e., composite sampling in this context will work well when there is a clear distinction between those units with and without the trait.

- Analytical results are available in a timely manner. (Some retesting schemes need the test results to be available sequentially; others do not. However, no composite sampling and retesting scheme will be practical if acquiring test results is a time-consuming process.)

### 11.1.3 Benefits

The benefit of composite sampling for this purpose is the cost savings that can be achieved. Substantial gains can be realized if the trait of interest is rare and the analytical costs are high relative to the costs of collecting, handling, maintaining, and compositing samples.

### 11.1.4 Limitations

The main limitations are those implied by the last 5 bullets of Section 11.1.2. Of particular concern is the condition when analyte presence/absence is the trait of interest, since mis-classification probabilities can become non-negligible due to dilution effects (which yield false acceptance results).

### 11.1.5 Implementation

Box 11-1 gives directions for use of composite sampling and retesting to classify units when the trait of interest is a *binary trait*. Various composite sampling and retesting schemes are available. These are listed in Table 11-1 (see Section 11.1.7 for examples). To determine an appropriate scheme, one first needs to determine if the test procedures allow the classification results to be immediately available (see second column of table). If so, the sequential types of schemes can be considered, but if not, one of the non-sequential schemes should be used. In general, the ordering of the schemes (rows) in Table 11-1 is from least to most complex and from least to most cost efficient. Table 11-1 (column 4) also refers to Tables 11-2, 11-3, and 11-4, which give optimal k values and relative costs for three of the schemes. The costs are based on the assumption that the positive units occur at random. However, if available information allows grouping the more similar units (with respect to the trait) into composites, then such grouping will reduce the expected number of tests and hence the study cost.

When the trait is based on whether a *continuous, non-negative measure exceeds or does not exceed a threshold* level $C$, then the fundamental difficulty is that a positive composite (a composite with level above $C/k$) does not necessarily imply that one or more of its component units is positive (has a level above $C$). In this case, the choice of $k$ not only considers the rarity of the trait but also the detection limit (DL) and how $C$ relates to the $DL$. In particular, $k$ must be less than $C/DL$, where $DL$ is a detection limit for which the likelihood of detection is high if concentrations at that level occur. As previously noted, the determination of (relative) cost is difficult in this situation, since it is not simply a

**Box 11-1. Generic Algorithm for use with the Various Schemes**

STEP 1: Based on prior knowledge, determine an upper bound on $p$, the proportion of the population having the trait of interest.

STEP 2: Determine the maximum number of aliquots that can be composited, based on the physical nature of the samples and their volumes and any concerns about dilution effects and the capability to combine and homogenize them adequately. Denote this maximum number of aliquots as $K$.

STEP 3: Decide on appropriate composite sampling and retesting scheme, based on whether classification results are available immediately (see Table 11-1) and whether a sequential scheme seems practical.

STEP 4: For the given scheme, determine if the optimal $k$ is less than $K$. If so, use the optimal k; if not, use $k=K$.

STEP 5: Estimate costs for the selected $k$ and anticipated prevalence, based on the number of units to be classified. (Formulae depend on the scheme.)

STEP 6: If the estimated cost exceeds the available resources, then consider cost-efficiency tradeoffs to arrive at useful $m$ and $k$ values. (Note that the restriction that k be less than or equal to $K$ may lead to situations in which an optimal combination of m and k is not possible.)

STEP 7: Implement the protocol. (Note that grouping units that are more likely to have the trait into a single composite will result in fewer analyses than the number that would occur by chance (random groupings), so that type of information, if available, should be used.) In forming each composite, thoroughly mix the material in each of the samples and obtain equal-volume aliquots from each one; then form the composite and homogenize the composite thoroughly prior to making the classification measurement.

function of $k$ and the trait's rarity but also depends on the distribution of contaminant over the site or process. Thus tables showing optimum $k$ values cannot be given for this case.

For this second case, if test results are not immediately available, one approach is simply to test composites versus a threshold $C/k$ and to retest all units in any positive composite. If test results are immediately available, another possible scheme for this case is the Dorfman-Sterrett Retesting scheme. This approach is similar to the curtailed exhaustive retesting scheme for binary traits (see Table 11-1); it relies heavily on the assumption of minimal measurement error. If $Y$ denotes the composite-sample result and $X_j$, $j$=1,2,...,$k$ denote potential individual-unit results, then the composite will be called

## Table 11-1. Identification of Composite Sampling and Retesting Schemes for Classifying Units Having a Rare Trait

| Name of Scheme | Sequential test results available immediately? | Description of Procedure | Table Providing Optimal k Values | Reference for Exact Algorithm |
|---|---|---|---|---|
| Exhaustive Retesting (Dorfman) | No | Each composite is tested; for each composite that tests positive, all individual samples are tested (at least one is positive). | 10-6 | Patil et al. (1996) based on Samuel (1978) |
| Curtailed (Dorfman) | Yes | Each composite is tested; for each composite that tests positive, the individual samples are tested sequentially; if k-1 have tested negative, then the $k^{th}$ unit is taken to be positive (without testing). | | Patil et al. (1996) |
| Sequential Retesting (Sterrett) | Yes | Each composite is tested; for each composite that tests positive, the individual samples are tested sequentially until a positive one is found; composites of the remaining samples are formed and tested in similarly. | 10-7 | Patil et al. (1996) based on Sterrett (1957). |
| Curtailed (Sterrett) | Yes | Like Sequential Retesting, except that if all but one of units from a positive composite have been tested and found to be negative, then testing of last unit is unnecessary (since it is known to be positive). | | Patil et al. (1996) |
| Binary Split Retesting | No | Each composite of $k$ units is tested; for each composite that tests positive, new sub-composites containing $k/2$ units are formed and tested; this splitting and retesting is continued until all units are classified. | 10-8 | Patil et al. (1996) based on Gill and Gottlieb (1974) |
| Curtailed Binary Split Retesting | Yes | Like the Binary Split Retesting scheme, except that only a single subcomposite of $k/2$ units (from a positive composite) is formed and tested; if it is found negative, a sub-subcomposite of $k/4$ units is formed and tested; this splitting and retesting is continued until all samples are classified. | | Patil et al. (1996) based on Gill and Gottlieb (1974) |
| Entropy-Based Retesting | Yes | Composites are formed sequentially; when a positive composite of k units is found, a single subcomposite containing $k/2$ units is formed; if it is positive, it is handled as in the curtailed binary split retesting scheme; but whenever a negative result occurs for the first subcomposite (or sub-subcomposite), the remaining units are grouped with those remaining to be classified and a new composite of $k$ units is formed. | | Patil et al. (1996) based on Hwang (1984) |

positive if $kY>C$. Note that $kY$ is expected to equal the sum of the $X$'s (since $Y$ is a physically averaged mean). Thus, if $kY<C$, then all of the $X$'s are less than $C$ and no further testing of the units in the composite is needed. Otherwise, testing continues sequentially until:

$kY - X_1 < C$ (which implies $X_2, X_3, ..., X_k$ are all less than C), or
$kY - X_1 - X_2 < C$ (which implies $X_3, X_4, ..., X_k$ are all less than C), or
$kY - X_1 - X_2 - X_3 < C$ (which implies $X_4, X_5, ..., X_k$ are all less than C), or
$kY - X_1 - X_2 - ... - X_{k-1} < C$ (which implies $X_k$ is less than C).

Actually, testing of the $k^{th}$ sample is unnecessary, so at most $k$ tests per composite are needed. The specific algorithm for this case is given by Patil et al. (1996).

A second approach, which also assumes that the test results are immediately available, is a Binary Split Retesting scheme. This scheme is similar to the curtailed binary split retesting scheme for binary traits; it also relies heavily on the assumption of no measurement error, since its testing-retesting strategy is based on same logic as for the prior scheme. The specific algorithm for this case is given by Patil et al. (1996).

**Table 11-2. Optimal Number of Samples per Composite for Exhaustive Retesting**

| Anticipated Prevalence (p) | Optimum k | Relative Cost* |
|---|---|---|
| 0.31 and above | 1 | 1 |
| 0.13, 0.14,..., 0.30 | 3 | 0.67, 0.70, ... , 0.99 |
| 0.07, 0.08,...,0.12 | 4 | 0.50, 0.53, ... , 0.65 |
| 0.05, 0.06 | 5 | 0.43, 0.47 |
| 0.03, 0.04 | 6 | 0.33, 0.38 |
| 0.02 | 8 | 0.27 |
| 0.01 | 11 | 0.20 |
| 0.005 | 15 | 0.14 |
| 0.001 | 32 | 0.06 |

* Based on random occurrence of positives. Equals $1+(1/k)+(1-p)^k$.
Source: Patil et al. (1996).

**Table 11-3. Optimal Number of Samples per Composite for Sequential Retesting**

| Anticipated Prevalence (p) | Optimum k | Relative Cost* |
|---|---|---|
| 0.31 and above | 1 | 1 |
| 0.22, 0.23,..., 0.30 | 3 | 0.843, 0.863,..., 0.992 |
| 0.13, 0.14,...,0.21 | 4 | 0.629, 0.654, ... , 0.822 |
| 0.09, 0.10, 0.11, 0.12 | 5 | 0.510, 0.541, 0.571, 0.600 |
| 0.06, 0.07, 0.08 | 6 | 0.406, 0.443, 0.478 |
| 0.05 | 7 | 0.367 |
| 0.04 | 8 | 0.324 |
| 0.03 | 9 | 0.276 |
| 0.02 | 11 | 0.221 |
| 0.01 | 15 | 0.152 |
| 0.005 | 21 | 0.106 |
| 0.001 | 45 | 0.046 |

* Based on random occurrence of positives. Equals $2 - q + (3q-q^2)/k - [1-q^{k+1}]/(kp)$ where $q=1-p$.
Source: Patil et al. (1996).

### 11.1.6 Relationship to Other Sampling Designs

Typically, when attempting to classify units, a finite population is of interest and one would use a census (for example, all wells in a region or drums in a field). If random sampling or grid sampling is employed, then inferences are restricted to the set of selected sampling units.

Judgment can be used to form composites if there is adequate information. If the goal is to classify all units, then statistical inference from a sample to a population is not relevant. Hence, if one has knowledge about which units are more homogeneous *with respect to the likelihood of having the trait* (as opposed to simply random choices), then compositing those units believed to be most alike will be more efficient than random groupings (if the information is correct). In this situation (i.e., a situation in which stratification or ranking might be used), fewer positive composites should occur than in a simple random sample situation and hence less retesting will be needed.

**Table 11-4. Optimal Values of *k* for Binary Split Retesting**

| Anticipated Prevalence (p) | Value of m (number of initial composites of size k) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **2** | **4** | **6** | **8** | **10** | **12** | **14** | **16** |
| >0.38 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.28 - 0.38 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 0.27 | 2 | 2 | 2,3 | 2 | 2 | 2 | 2 | 2 |
| 0.26 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 |
| 0.25 | 2 | 2 | 3 | 2,3 | 2,3 | 2,3 | 2,3 | 2 |
| 0.24 | 2 | 2,4 | 3 | 3 | 3 | 3 | 3 | 2,3 |
| 0.20 - 0.23 | 2 | 4 | 3 | 3 | 3 | 3 | 3 | 3 |
| 0.19 | 2 | 4 | 3 | 3,4 | 3,4 | 3,4 | 3 | 3,4 |
| 0.18 | 2 | 4 | 3 | 4 | 4 | 4 | 3,4 | 4 |
| 0.17 | 2 | 4 | 3,4 | 4 | 4,5 | 4 | 4 | 4 |
| 0.16 | 2 | 4 | 4,6 | 4 | 5 | 4 | 4,5 | 4 |
| 0.15 | 2 | 4 | 6 | 4 | 5 | 4 | 5 | 4 |
| 0.14 | 2 | 4 | 6 | 4 | 5 | 4,5 | 5 | 4 |
| 0.13 | 2 | 4 | 6 | 4 | 5 | 5,7 | 5 | 4,7 |
| 0.12 | 2 | 4 | 6 | 4,8 | 5 | 7 | 5,7 | 7 |
| 0.11 | 2 | 4 | 6 | 8 | 5 | 7 | 7 | 7,8 |
| 0.10 | 2 | 4 | 6 | 8 | 5,10 | 7 | 7 | 8,11 |
| 0.09 | 2 | 4 | 6 | 8 | 10 | 7 | 7 | 11 |
| 0.08 | 2 | 4 | 6 | 8 | 10 | 7,12 | 7 | 11,16 |
| 0.00 - 0.07 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 |

Source: Patil et al. (1996).

### 11.1.7 Examples

Figure 11-1 illustrates how several of the composite sampling and retesting schemes would perform for a case in which classification is based on a binary trait. The example involves a situation in which 32 units (numbered 1 to 32) are to be classified and for which three particular units have the trait— namely, units 6, 19, and 20. As an illustration of how to interpret Figure 11-1, consider the binary split algorithm. At the first stage, aliquots are used to form four composites of eight units each. The first and third composites test positive because they contain unit 6 and units 19 and 20, respectively, while the second and fourth composites test negative. Further testing is needed for the units in the first and third composites, so at Stage 2 four new composites are formed, each containing four units. Since two of these composites test positive, Stage 3 involves forming four new composites of two units each, two of which test positive. Finally, at Stage 4, aliquots from four individual samples are tested and the three positive units are identified. The overall process involved a total of 16 tests for this example, as compared to 32 if the individual units had been tested. The figure shows that the various composite sampling and retesting schemes needed from 13 to 20 tests for this example. As illustrated by this example, the usual situation would involve classification of all units in some given population—for instance, classification of all drinking water wells in a given region to determine those that have animal waste contamination.

An application of the binary split algorithm for testing drinking water wells for pesticide contamination is described by Natarajan and Rajagopal (1993). Spiked samples were also used to assess the performance of the procedure.

## 11.2 COMPOSITE SAMPLING AND RETESTING FOR IDENTIFYING EXTREME SAMPLING UNITS

### 11.2.1 Overview

All units in a given population might be exhaustively tested if the interest is in identifying the maximum value (for example, concentration of contaminant $X$) that occurs among those units (for example, $n$ drums or $n$ spatial grid points). However, under certain circumstances, composite sampling coupled with the retesting of some units in one or more of the composites with the highest levels can reveal the unit with the maximum level—and at a substantially reduced cost. Similar composite sampling and retesting strategies can be used to identify the second highest unit, the third highest unit, etc.—i.e., to determine the upper percentiles of the population. This use of compositing has recently received some theoretical attention; real examples, however, do not appear to exist.

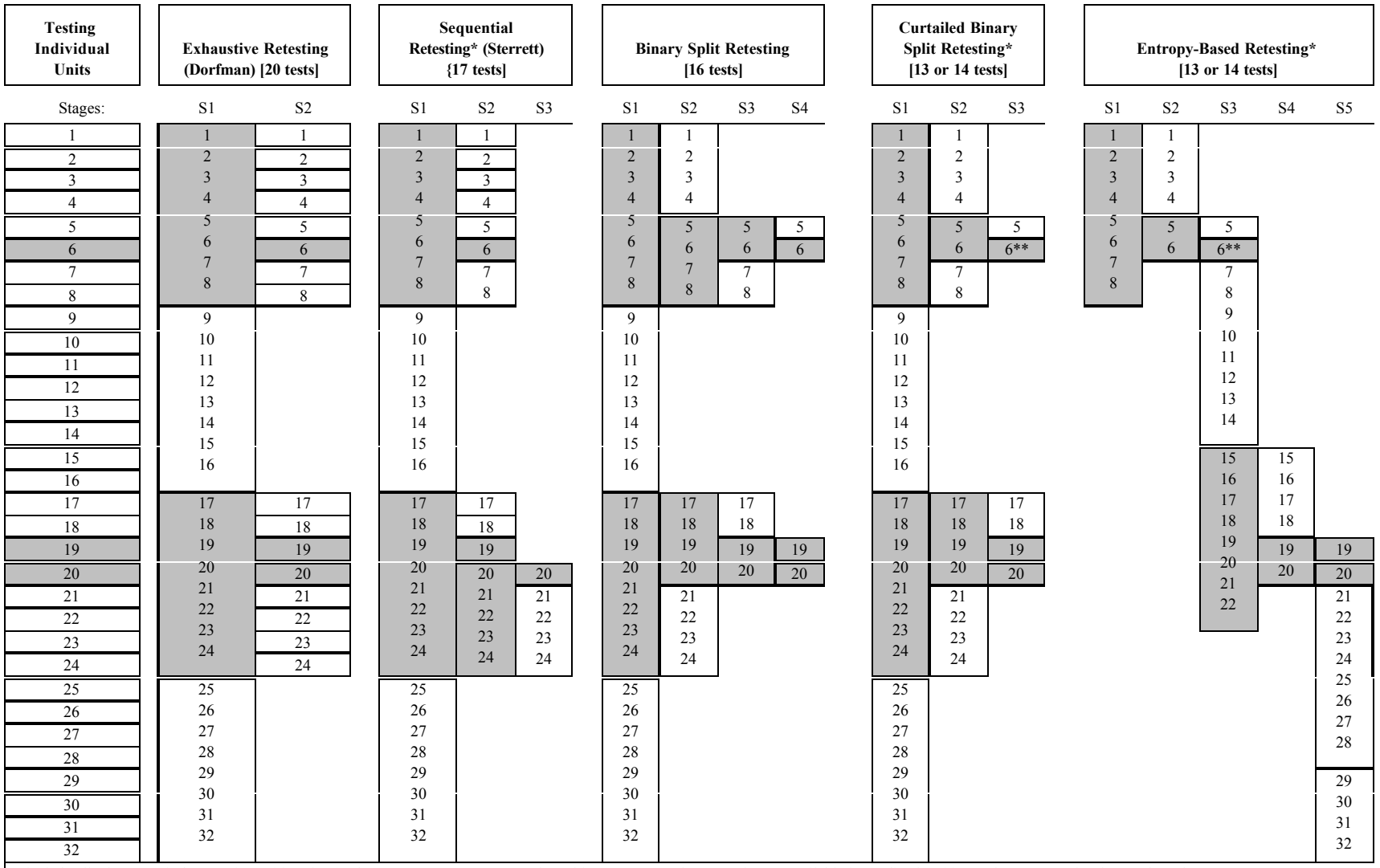


**Figure 11-1. Illustration of Retesting Schemes for Classifying Units When 3 of 32 Units are Positive**

### 11.2.2 Application

It will generally be appropriate to use composite sampling and retesting for determining the unit with maximum value (and the value itself) when the following conditions hold:

1. There is a predefined set of units that are to be evaluated.
2. Analytical costs are high relative to costs associated with sampling, sample acquisition, sample handling, and compositing. (Otherwise, composite sampling and retesting will not be cost-effective.)
3. Representative aliquots from the individual units can be obtained.
4. There are no practical difficulties in forming appropriate composites from aliquots (for example, aliquots can be adequately homogenized).
5. Compositing will not affect the sample integrity.
6. Retesting of units is feasible. In particular, the identity of the units can be maintained, and samples can be adequately preserved throughout all the potential stages of testing and retesting.
7. Measurement error, relative to the range of $X$ concentrations, must be minimal. This assumption is necessary because the tests on individual units or on composites of a given size have to be properly ordered.
8. Analytical results are available in a timely manner. (Otherwise, the retesting schemes will be impractical.)

### 11.2.3 Benefits

Costs can be substantially less than using exhaustive testing.

### 11.2.4 Limitations

The main limitations are those implied by conditions (3) through (8) in Section 11.1.4, especially condition (7), the assumption of negligible measurement error.

### 11.2.5 Implementation

Several methods have been proposed for using composite sampling and retesting to identify a maximum. These include the following:

- Casey et al. (1985) proposed a simple method for predicting a maximum in which searching for the maximum is restricted to the composites having the highest levels. The simplest strategy would involve retesting only those units in the composite that had the highest level; this approach obviously cannot guarantee that one finds the unit with the maximum level.

- Gore and Patil (1994) proposed a *sweep-out method for identifying a maximum*. This method is based on the idea that if the composite with the highest observed value has a level that is "sufficiently" bigger than the level for the next highest composite, then one can be assured (apart from possible measurement error concerns) that the unit with highest level is in the composite with the highest level; hence retesting of the individual samples in that composite will reveal the unit with the maximum (with certainty, if there is no measurement error). However, if the "gap" in observed levels for these two composites is not "sufficiently" large, then retesting of units in some other composites may be needed to identify the one with the maximum level. The algorithm relies on the fact the maximum individual value within a composite consisting of $k$ units must fall in the interval $[Y, kY]$, where $Y$ is the value observed for the composite. This property assumes no measurement error is involved; it is true because the maximum is always larger than the average, $Y$, and smaller than the total, $kY$. Thus if the units in the composite with the highest $Y$ are each individually tested and the maximum value is found to be $Z$, then the composite measurements for each of the other composites can be compared with $Z$ to determine if they might contain the maximal unit: only composites where $kY > Z$ can contain the maximal unit and hence only the units within those composites need further testing. The explicit algorithm for this method is given by Patil et al. (1996).

- An extension of the *sweep-out method for identifying a maximum* is a *sweep-out method for identifying upper percentiles*, which is described by Patil et al. (1996).

- Two other methods, which are also extensions of the *sweep-out method for identifying a maximum* and which are described by Gore et al. (1996), are a *locally-sequential sweep-out method* and *globally-sequential sweep-out method* for identifying a maximum.

These references mention some examples wherein these techniques may be applied.

### 11.2.6 Relationship to Other Sampling Designs

Most commonly, there is a finite population of interest. A grid sample might be viewed as a surrogate "census" if interest is confined to the "sample." With a maximum being the parameter of interest, no statistical inference beyond the sample (whatever it may be) is generally warranted.

## GLOSSARY OF TERMS

**acceptance criteria** - specific limits placed on the characteristic of an item, process, or service.

**action level** - the numerical value that causes a decision maker to choose or accept one of the alternative actions (for example compliance or noncompliance) to the no action alternative. It may be a regulatory threshold standard, such as a maximum contaminant level for drinking water; a risk-based concentration level; a technology limitation; or reference-based standard. Note that the action level defined here is the planning phase of a data collection activity; it is not calculated from the sampling data.

**adaptive sampling** - selecting an initial probability based sample followed by additional sampling that is then based on observed results, where follow-up samples are concentrated in areas of elevated levels of the feature of interest. Adaptive sampling is useful for locating wildlife concentrations, veins of precious metals, or identifying areas of environmental contamination, allowing the sampling team to determine regions of elevated levels of interest while maintaining the ability to estimate mean and variance. Disadvantages of adaptive sampling include the ability to only obtain biased estimates for population mean and variance, that random sample sizes are generally larger, and that iterative sampling may be time consuming.

**auxiliary variable** - a variable providing information helpful to developing the sampling design.

**bias -** the systematic or persistent distortion of a measurement process that causes errors in one direction (the expected sample measurement value differs from the sample's true value).

**binary trait** - a characteristic that can only have two possible values.

**boundaries** - the spatial and temporal conditions and practical constraints under which environmental data are collected. Boundaries specify the area of volume (spatial boundary) and the time period (temporal boundary) to which decisions will apply**.**

**coefficient of variation (CV)** - a unit-free measure of variability. The CV can be thought of as the standard deviation expressed as a percentage of the mean.

**composite sampling** - sampling method used where several samples are physically mixed into a larger composite sample. The entire composite sample may be measured for desired information, or one or more random sub-samples may be measured individually. In general, individual samples which are composited must be the same size or volume and the composite sample must be completely mixed. Composite sampling can be useful for estimating mean concentration of a substance, and if appropriate, compositing can result in substantial savings where the cost of analyzing individual samples is high.

**composite sampling protocol** - a description of the scheme for forming and processing (mixing and homogenizing) composites. This sampling protocol will indicate whether entire samples or aliquots are to be combined, the number of groups of units to be formed ($m$), the number of units per group ($k$), which units form each group, and the amount of material from each unit to be used in forming the composite sample.

**conceptual model** - a description of the expected source of the contaminant and the size and breadth of the area of concern, including relevant fate and transport pathways and potential exposure pathways.

**confidence interval** - an interval estimate of a population parameter with a known probability that the population value of interest will be included in the interval. For example, a 95% confidence interval estimate of a population mean is an interval that will contain the true value of the mean in 95% of all samples that could be selected with a given sampling design.

**decision error** - an error which occurs when data misleads a site manager into choosing the wrong response action, in the sense that a different action would have been taken if the site manager had access to unlimited "perfect data" or absolute truth. In a statistical test, decision errors are labeled as false rejection or false acceptance depending on the concerns of the decision maker and the baseline condition chosen.

**defensible** - the ability to withstand any reasonable challenge related to the veracity or integrity of project and laboratory documents and derived data.

**detection limit (DL)** - a measure of the capability of an analytical method of distinguished samples that do not contain a specific analyte from a sample that contains a low concentration of the analyte; the level of target analyte that can be determined to be different from zero by a single measurement at a stated level of probability DLs are analyte- and matrix-specific and may be laboratory dependent.

**distribution** - (1) the concentration of an environmental contaminant at a point over time, over an area, or within a volume; (2) a probability function (density function, mass function, distribution function) used to describe a set of observations or a population from which the observations are generated.

**double sampling** - measuring a characteristic of a sample that is positively correlated with the outcome of interest, with a subsample of those units selected for measuring the primary outcome. The relationship between the two variables is used to generate the data needed for the primary outcome variable.

**environmental data** - any measurements or information that describe environmental processes, location, or conditions; ecological or health effects consequences; or the performance of environmental technologies. For EPA, environmental data include information collected directly from measurements, produced from models, and complied from others sources such as data bases or the literature.

**estimate** - a characteristic from the sample from which inferences on parameters can be made.

**false acceptance decision error** - the error that occurs when a decision maker accepts the baseline condition when it is actually false. Statisticians usually refer to the limit on the possibility of a false acceptance decision error as beta ($) and it is related to the power of the statistical test used in decision making. May also be referred to as a false negative decision error. Also known as a Type II Error.

**false rejection decision error** - the error that occurs when a decision maker rejects the baseline condition when it is actually true. Statisticians usually refer to the limit on the possibility of a false acceptance decision error as alpha (a). Also known as a Type I Error.

**grid sampling** - A method for determining the location of samples where sample locations are located at the nodes of a geometrical grid pattern (for example, square, rectangle, triangle, hexagon).

**hypothesis test** - a statistical procedure for determining if a sample provides sufficient evidence to reject one statement regarding the population of interest (the null hypothesis) in favor of an alternative statement (the alternative hypothesis). The null hypothesis is considered the "baseline" condition and will be rejected in favor of the alternative hypothesis only when there is overwhelming evidence the null cannot be true.

**judgmental sampling** - use of professional judgment to select sampling locations.

**mean (sample)** - the average of the sample values.

**mean (population)** - the average of the population values.

**measurement protocol** - a specified procedure for making observations or performing analyses to determine the characteristics of interest for each sampling unit. Measurement protocols include the procedures for collecting a physical sample, handling and preparing the physical sample, and applying an analytical method to obtain a result.

**population** - the total collection of objects or people to be studied and from which a sample is to be taken.

**population parameter** - a characteristic defined in terms of all units in a population.

**probability-based sampling** - a method of selecting samples such that the probability of being included in the sample is known for every unit on the sampling frame.

**ranked set sampling** - a field sampling design wherein expert judgment or an auxiliary measurement method is used in combination with simple random sampling to determine which locations in the field

should be collected to be measured by the method of choice. An expert or auxiliary measurement is used to rank (order) field locations (selected using simple random sampling) with respect to the variable of interest to determine which locations to sample.

**representativeness** - a measure of the degree to which data accurately and precisely represent characteristics of a population, parameter variations at a sampling point, a process condition, or an environmental condition. Representativeness is also the correspondence between the analytical result and the actual environmental quality or condition experienced by a contaminant receptor.

**sample** - a set of units or elements selected from a larger population, typically to be observed for making inferences regarding that population.

**sample size** - the number of sample units to be collected.

**sample support** - the portion of the sampling unit that is extracted in the field and that is subjected to the measurement protocol-- the area or volume that a single sample is supposed to represent.

**sampled population** - the set of units or elements from which a sample was selected, i.e., the units that had a chance of being included in the sample.

**sampling design** - a description of the sample collection plan that specifies the number, type, and location (spatial and/or temporal) of sampling units to be selected for measurement.

**sampling frame** - the list from which a sample of units or elements is selected.

**sampling unit**- the members of a population that may be selected for sampling.

**simple random sampling** - method of sampling where samples are collected at random times or locations throughout the sampling period or study area.

**standard deviation** - a measure of the dispersion or imprecision of a sample or population distribution as expressed as the square root of the variance and has the same unit of measurement as the mean. The standard deviation is calculated as the square root of the variance.

**standard error** - the standard error of a sample statistic, 2, (for example, a sample mean or proportion) is the standard deviation of the values of that statistic over repeated samples using the same sampling design (for example, stratified simple random sampling) and the same sample size, $n$.

**statistic** - a quantity calculated from the values in a sample (for example, the sample mean or sample variance)**.**

**stratified sampling** - sampling method where a population is divided into non-overlapping sub-populations called strata and sampling locations are selected independently within each strata using some sampling design. Stratified sampling is used to make inferences about individual strata that are inherently more homogeneous than the entire heterogeneous population**.**

**systematic sampling** - A method for determining the location of samples in which only one unit (in space or time) is selected randomly. This randomly selected unit establishes the starting place of a systematic pattern (for example, every 3 days, every 5$^{th}$ unit, every unit at a node on a grid design) that is repeated throughout the population.

**target population** - the set of all units or elements (for example, barrels of waste or points in time and/or space) about which a sample is intended to draw conclusions.

**Type I error** - the statistical term for false rejection decision error.

**Type II error** - the statistical term for false acceptance decision error.

**variability** - observed difference attributable to heterogeneity or diversity in a population. Sources of variability are the results of natural random processes and stem from environmental differences among the elements of the population. Variability is not usually reducible by further measurement but can be better estimated by increasing sampling.

**variance (sample)** - a measure of the spread of the sample. A larger variance means a larger spread. This is an estimate of population variance.

**variance (population)** - a measure of the spread of the population. A larger variance means a larger spread. In most cases, the population variance is only a theoretical quantity-- it cannot be known with certainty because all possible data points in the population cannot be measured.

## BIBLIOGRAPHY


ANSI/ASQC 1994. *Specifications and Guidelines for Environmental Data Collection and Environmental Technology Programs (E4)*. American National Standards Institute (ANSI) and American Society for Quality Control (now American Society for Quality).

ASTM D6232-00 (2000). *Standard Guide for Selection of Sampling Equipment for Waste and Contaminated Media Data Collection Activities*. American Society for Testing and Materials, West Conshohocken, PA.

Barabesi, L. 1998. The computation of the distribution of the sign test statistic for ranked-set sampling, *Communications in Statistics-Simulation* 27(3):833-842.

Bohn, L.L., and D.A. Wolfe. 1992. Nonparametric two-sample procedures for ranked-set samples data. *Journal of the American Statistical Association* 87:552-561.

Bohn, L.L., and D.A. Wolfe. 1994. The effect of imperfect judgment ranking on properties of procedures based on the ranked-set samples analog of the Mann-Whitney-Wilcoxon statistic. *Journal of the American Statistical Association* 89:168-176.

Boswell, M.T., S.D. Gore, G. Lovison and G.P. Patil. 1992. *Annotated bibliography of composite sampling*, Technical Report Number 92-0802, Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA.

Cailas, M.D., R.G. Kerzee, E.K. Mensah, K.G. Croke, and R. R. Swager. 1995. A proposed methodology for an accurate estimation of the total amount of materials recycled. Resources. *Conservation and Recycling* 15, 123-131.

Casey, D.B., P.N. Nemetz, and D. Uyeno. 1985. Efficient search procedures for extreme pollutant values. *Environmental Monitoring and Assessment* 5:165-176.

Christman, M.C. 1997. Efficiency of adaptive sampling designs for spatially clustered populations. *Environmetrics* 8:145-166.

Christman, M.C. 2000. A review of quadrant-based sampling of rare, geographically clustered populations. *Journal of Agricultural, Biological, and Environmental Statistics* 5(2):168-201.

Cochran, W.G., and G.M. Cox. 1957. *Experimental Designs*. John Wiley & Sons, New York.

Cochran, W.G. 1963. *Sampling Techniques*, 2nd ed. John Wiley & Sons, New York.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillside, NJ.

Cox, D.D., Cox, L.H, and K.B. Ensor. 1995. Spatial Sampling and the Environment. *National Institute of Statistical Sciences Technical Report #38*, Research Triangle Park, NC.

Cressie, N.A. 1993. *Statistics for Spatial Data.* John Wiley & Sons, New York.

Dalenius, T., and J.L. Hodges, Jr. 1959. Minimum variance stratification. *Journal of the American Statistical Association* 54:88-101.

Davidson, J.R. 1995. *ELIPGRID - PC's User's Manual*. ORNL/TM-13103, Oak Ridge National Laboratory-Grand Junction, Grand Junction, CO. Software available for download at: http://www.epa.gov/quality

Dorfman, R. 1943. The detection of defective members of large populations. *Annals of Mathematical Statistics* 14:436-440.

Elder, R.S., W.O. Thomson, and R.H. Myers. 1980. Properties of composite sampling procedures. *Technometrics* 22(2):179-186.

Entz, T. and C. Chang. 1991. Evaluation of soil sampling schemes for geostatistical analyses: A case study for the soil bulk density. *Canadian Journal of Soil Science* 17:165-176.

Garner, F.C., M.A. Stapanian, and L.R. Williams. 1988. Composite sampling for environmental monitoring. Chapter 25 in *Principles of Environmental Sampling,* L.H. Keith, ed. American Chemical Society, Washington DC.

Garner, F.C., M.A. Stapanian, E.A. Yfantis, and L.R. Williams. 1989. Probability estimation with sample compositing techniques. *Journal of Official Statistics* 5:365-374.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York.

Gilbert, R.O. 1995. Ranked Set Sampling. *DQO Statistics Bulletin, Statistical Methods for the Data Quality Objectives Process*, Volume 1, Number 1. PNL-SA-26377. Pacific Northwest National Laboratory, Richland, WA.

Gill, A., and Gottlieb. 1974. The identification of a set by successive intersections. *Information and Control* 24:20-35.

Ginevan, M. 2001. Using Statistics in Health and Environmental Risk Assessments. A Practical Guide to Understanding, Managing, and Reviewing Environmental Risk Assessment Reports, S.L. Benjamin and D.A. Bullock, eds., Lewis, New York.

Gore, S.D., and G.P. Patil. 1994. Identifying extremely large values using composite sample data. *Environmental and Ecological Statistics* 1:227-245.

Gore, S.D., G.P. Patil, and C. Taillie. 1996. Identification of the largest individual sample value using composite sample data and certain modifications to the sweep-out method. *Environmental and Ecological Statistics* 3:219-234.

Hansen, M.M., and W.N. Hurwitz. 1943. On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 14:333-362.

Hettmansperger, T.P. 1995. The ranked-set sample sign test, *Nonparametric Statistics* 4:263-270.

Horvitz, D.G. and D.J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47:663-685.

Isaaks, E.H. and Srivastava, R.M.. 1989. *Introduction to Applied Geostatistics*, Oxford University Press, New York.

Johnson, G.D., B.D. Nussbaum, G.P. Patil and N.P. Ross. 1996. Designing cost-effective environmental sampling using concomitant information, *Chance* 9(1):4-11.

Kaur, A., G.P. Patil and C. Taillie. 1995. *Unequal Allocation Models for Ranked Set Sampling with Skew Distributions*, Technical Report Number 94-0930, Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA

Koti, K.M. and G.J. Babu. 1996. Sign test for ranked-set sampling, *Communications in Statistics-Theory and Methods* 25(7):1617-1630.

Kraft, K.M, D.H. Johnson, J. M. Samuelson, and S.H. Allen. 1995. Using Known Populations of Pronghorn to Evaluate Sampling Plans and Estimators. *Journal of Wildlife Management* 59(1), 129-137.

Lancaster, V., and S. Keller-McNulty. 1998. Composite sampling, part I. *Environmental Testing and Analysis* 7(4):15ff.

Li, F., and J. Chaplin. 1998. Evaluation of large field sampling methods for crop residue coverage measurement. *Transactions of the American Society of Agricultural Engineers* 41(3):645-651.

Midwest Plan Service. 1992. *Conservation Tillage Systems and Management*. MWPS-45. MidWest Plan Service. Iowa State University, Ames, IA.

McIntyre, G.A. 1952. A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research* 3:385-390.

Mode, N.A., L. C. Conquest, and D. A. Marker. 1999. Ranked set sampling for ecological research: accounting for the total costs of sampling. *Environmetrics* 10:179-194.

Muttlak, H.A. 1995. Parameters estimation in a simple linear regression using rank set sampling, *Biometrical Journal* 37:799-810.

Muttlak, H.A. 1996. Estimation of parameters for one-way layout with rank set sampling, *Biometrical Journal* 38:507-515.

Myers, J.C. 1997. *Geostatistical Error Management: Quantifying Uncertainty for Environmental Sampling and Mapping*. Van Nostrand Reinhold, New York.

Natarajan, U. and R. Rajagopal. 1993. Sample compositing. *Environmental Testing and Analysis* 2(3):54ff.

Nussbaum, B.D. and B.K. Sinha. 1997. Cost effective gasoline sampling using ranked set sampling. *Proceedings of the Section on Statistics and the Environment*. American Statistical Association, pp. 83-87. American Statistical Association, Alexandria, VA.

Patil, G.P., A.K. Sinha and C. Taillie. 1994. *Ranked set sampling, Handbook of Statistics 12, Environmental Statistics* pp. 167-200, (G.P. Patil and C.R. Rao, editors), North-Holland, New York, NY

Patil, G.P., S.D. Gore, and G.D. Johnson. 1996. *EPA Observational Economy Series, Volume 3: Manual on Statistical Design and Analysis with Composite Samples* (Draft). Technical Report No. 96-0501, Center for Statistical Ecology and Environmental Statistics Pennsylvania State University.

Perez, A., and J.J. Lefante. 1997. Sample size determination and the effect of censoring when estimating the arithmetic mean of a lognormal distribution. *Communications in Statistics, Theory and Methods* 26 (11):2779-2801.

Pitard, F.F.. 1993. *Pierre Gy's Sampling theory and Sampling Practice*, 2nd ed. CRC Press, Boca Raton, FL.

Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. 1992. *Numerical Recipes in Fortran 77: The Art of Scientific Computing.* Cambridge University Press, Cambridge, MA.

Samawi, H.M. and H.A. Muttlak. 1996. Estimation of ratio using rank set sampling, *Biometrical Journal* 38:(6):753-764.

Seber, G.A.F. and S.K. Thompson. 1994. Environmental Adaptive Sampling. *Handbook of Statistics, Vol. 12 (Environmental Sampling)*. G.P. Patil and C.R. Rao, Eds. Elsevier Science B.V., New York.

Starks, T.H. 1986. Determination of support in soil sampling. *Mathematical Geology* 18(6):529-537.

Stehman, S.V. and Overton, S. (1994), Comparison of Variance Estimators of the Horvitz-Thompson Estimator for Randomized Variable Probability Systematic Sampling, *Journal of the American Statistical Association* 89, 30-43.

Sterrett, A. 1957. On the detection of defective members of large populations. *Annals of Mathematical Statistics*, 28: 1033-1036.

Stokes, S.L. and T.W. Sager. 1988. Characterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association* 83:374-381.

Thompson, S. K. 1990. Adaptive Cluster Sampling. *Journal of the American Statistical Association* 85:412.

Thompson, S. K. 1992. *Sampling*. John Wiley & Sons, New York.

Thompson, S.K. and G.A.F. Seber. 1996. *Adaptive Sampling*. John Wiley & Sons, New York.

U. S. Environmental Protection Agency. 1989. *Methods for Evaluating the Attainment of Cleanup Standards*, Vol. 1. Soils and Solid Media. PB89-234959. Washington, DC.

U.S. Environmental Protection Agency. 1992. *Statistical Methods for Evaluating the Attainment of Cleanup Standards, Volume 3: Reference-Based Standards for Soils and Solid Media*. NTIS # PB-94-176-831. Office of Policy Planning and Evaluation, Washington, DC.

U.S. Environmental Protection Agency. 1996a. *Soil Screening Guidance: User's Guide.* EPA/540/R-96/0180. Office of Solid Waste and Emergency Response, Washington, DC.

U.S. Environmental Protection Agency. 1996b. *Geostatistical Sampling and Evaluation Guidance for Soils and Solid Media, Review Draft*, Office of Solid Waste, Washington, DC.

U.S. Environmental Protection Agency. 1998a. *Quality Assurance Guidance for Conducting Brownfields Site Assessments.* EPA/540/R-98/038. Office of Solid Waste and Emergency Response, Washington, DC.

U.S. Environmental Protection Agency. 1998b. *Guidance for Quality Assurance Project Plans (QA/G-5).* EPA/600/R-98/018. Office of Research and Development, Washington, DC.

U.S. Environmental Protection Agency. 2000a. *Guidance for Data Quality Assessment: Practical Methods for Data Analysis (QA/G-9).* EPA/600/R-96/084. Office of Research and Development, Washington, DC.

U.S. Environmental Protection Agency. 2000b. *Guidance for the Data Quality Objectives Process (QA/G-4).* EPA/600/R-96/055. Office of Environmental Information, Washington, DC.

U.S. Environmental Protection Agency. 2000c. *EPA Quality Manual for Environmental Programs*, EPA Manual 5360 A1. Washington, DC.

U.S. Environmental Protection Agency. 2001. Peer Review Draft. *Guidance for Data Quality Indicators (QA/G-5i).* Office of Environmental Information. Washington, DC.

Wang, X.J. and F. Qi. 1998. The effects of sampling design on spatial structure analysis of contaminated soil. *The Science of the Total Environment* 224, 29-41.

Wolter, K.M. 1984. An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association* 79, 781-790.

Yfantis, A.A., G.T. Flatman, and J.V. Behar. 1987. Efficiency of kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology* 19, 183-205.