An Overview of Methods for Evaluating the Attainment of Cleanup Standards for Soils, Solid Media, and Groundwater, EPA Volumes 1, 2, and 3

R. O. Gilbert
T. LeGore
R. F. O'Brien

January 1996

Prepared for the
U.S. Environmental Protection Agency
under Contract DE-AC06-76RLO 1830

Pacific Northwest National Laboratory
Operated for the U.S. Department of Energy
by Battelle

ACKNOWLEDGMENTS

This report was prepared by Richard O. Gilbert and Robert F. O'Brien of the Pacific Northwest National Laboratory (PNNL) and by Tim LeGore of Boeing Computer Services Richland for the U.S. Environmental Protection Agency (EPA). Barry Nussbaum and Ron Shafer of EPA, Office of Strategic Planning and Environmental Data, provided EPA project management and peer review. Derrick Bates of PNNL also peer reviewed draft chapters of this report and Frank Ryan of PNNL provided editorial review and assistance. Questions regarding this publication should be directed to Barry Nussbaum, Office of Strategic Planning and Environmental Data, U.S. Environmental Protection Agency, 401 M Street SW, Washington, DC 20460; Phone (202) 260-1493.

FOREWORD

The cleanup of a contaminated site is not complete until a defensible data-driven decision is made that residual concentrations of hazardous chemicals at the site are less than required by all applicable cleanup standards and guidelines. Statistical tests are important aids for making such decisions. They provide a formal and objective procedure for making quantitative decisions while taking into account the total variability among the measurements. They also control the probability of making incorrect decisions to acceptable levels, specified a priori.

The use of selected statistical tests to evaluate the attainment of risk-based or reference-based cleanup standards for soil, solid media or groundwater is described in three reports published by the U.S. Environmental Protection Agency (EPA). The purpose of this document is to provide an executive summary and overview of these three volumes for use by EPA Headquarters staff, EPA regional remedial project managers, potentially responsible parties for Superfund sites, the staff of State environmental protection agencies, and contractors for these groups. The primary goal of this overview is to provide the reader with an understanding of why these volumes are useful and how to use them.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION AND OVERVIEW

Remedial actions at a contaminated site are not complete until the attainment of cleanup standards applicable to the site has been verified. Statistical tests, if properly selected and used, are a powerful tool for verifying the attainment of standards. They provide an important input to quantitative, data-driven decision making. Moreover, they provide a method for formally and objectively handling variability among environmental measurements in the decision making process. Although it is impossible to assure that an incorrect decision will *never* be made, statistical tests provide a structure for limiting the decision error rates to acceptable low levels that have been specified by decision makers.

The U.S. Environmental Protection Agency (EPA) has published threee reports that describe and illustrate how to use statistical tests to evaluate the attainment of certain cleanup standards for soil, solid media and groundwater (USEPA 1989, 1992 and 1994b). These reports are henceforth denoted as Volumes 1, 2 and 3, respectively. The purpose of this document is to provide an executive summary and overview of these reports to facilitate their use by EPA managers and the staff of State environmental protection agencies, contractors and other interested government agencies.

Table 1.1 provides a summary of some of the key differences and similarities in the three volumes that are discussed in detail in Chapter 2.0. We see from the table that:

- Volume 1 applies to risk-based standards for soil and solid media
- Volume 2 applies to risk-based standards for groundwater
- Volume 3 applies to reference-based or background-based standards for soil and solid media

The general testing process used in Volumes 1, 2 and 3 is illustrated in Figure 1.1. The first step is to use the Data Quality Objectives (DQO) process to plan the study and specify the quality and quantity of data needed (the data quality objectives). After samples are collected and measured for contaminants, the Data Quality Assessment (DQA) process is used to determine if the quality and quantity of the data obtained meet the specified DQOs. If not, data of higher quality are needed. Once data of the required quality are obtained, the statistical test (or tests) is performed. If the test indicates the cleanup standard has been attained, a report is written and a briefing is given that describes the testing approach and rationale, the test results, and future activities that are required to assure continuing compliance with standards.

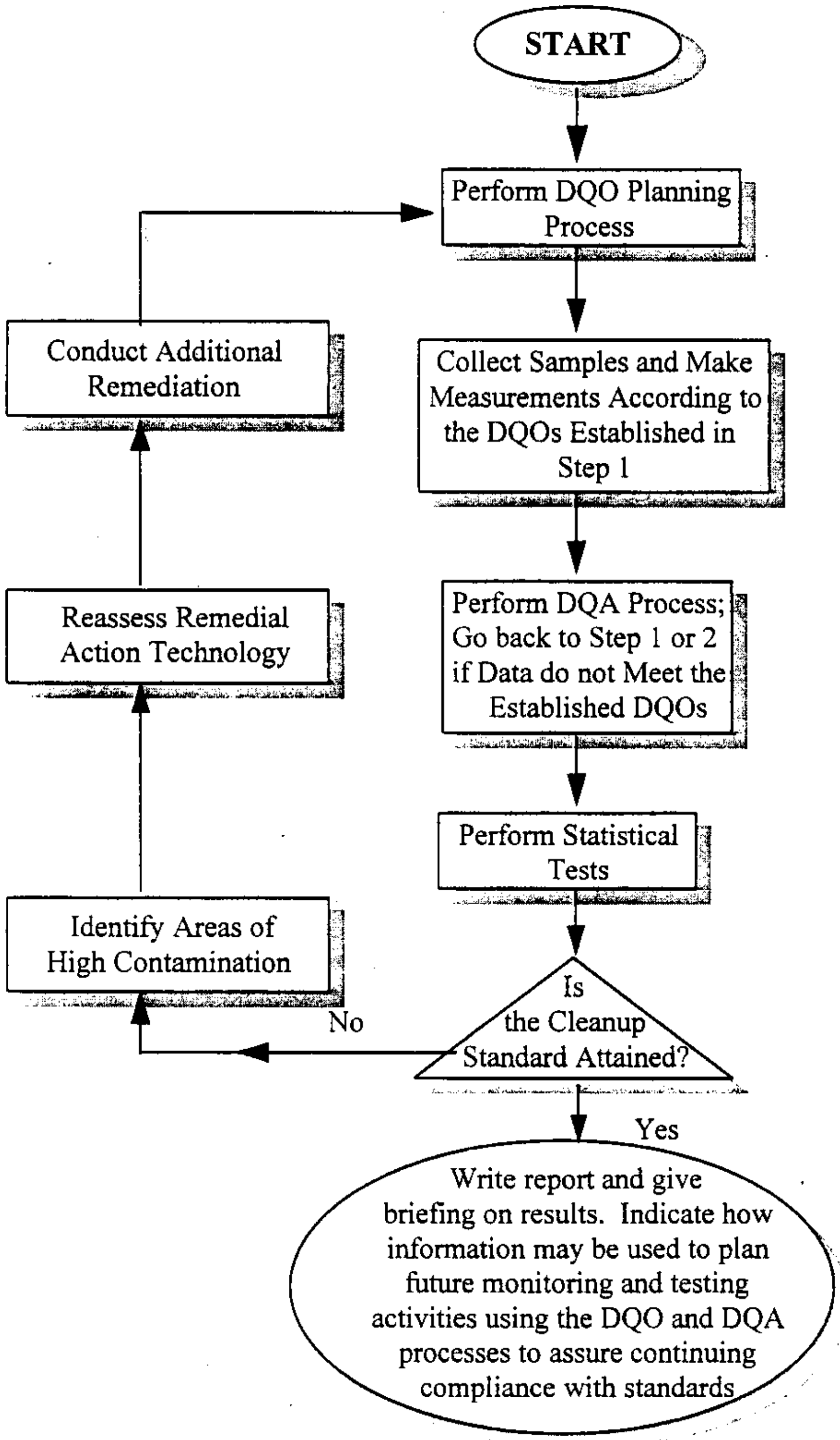
If the test indicates the standard has not been attained, then the areas of high contamination should be identified. This may require additional sampling. Then, following a reassessment of the remedial action technology to determine whether it should be improved, additional remediation is conducted. This is followed by again applying the DQO and DQA processes to assure that appropriate data are obtained to conduct statistical tests to evaluate the attainment of standards for the newly remediated site.

TABLE 1.1 Summary of Three EPA Reports (USEPA 1989, 1992, 1994b; Volumes 1, 2 and 3, respectively) that Provide Statistical Tests for Evaluating the Attainment of Cleanup Standards

Category	Volume 1	Volume 2	Volume 3
Environmental Media	Soil and solid media	Groundwater in wells	Soil and solid media
Cleanup Standard	Risk-based	Risk-based	Reference-based or background-based
Planning	DQO Process	DQO Process	DQO Process
Target Population	Area of land for which a decision is needed	Water at well locations, not water in the entire aquifer	Area of land for which a decision is needed
Hypotheses Used	H_0: Contaminated H_a: Clean	H_0: Contaminated H_a: Clean	H_0: Clean H_a: Contaminated
Sampling Designs Discussed	Simple random Stratified random Systematic Composite Sequential	Simple random Stratified random Systematic Sequential	Systematic Unaligned grid pattern Composite
How Non-Detects are Handled	Non-detects set equal to the detection limit	Non-detects set equal to the detection limit	All non-detects must be less than the smallest detected value in the data set
Statistical Tests Described	UCL[a] on site mean UCL on an upper percentile of the site data distribution UCL on the proportion of the site that is greater than the standard Number of measurements exceeding the standard	UCL on well mean UCL on an upper percentile of the site data distribution UCL on the proportion of the site that is greater than the standard	Wilcoxon Rank Sum Quantile Hot measurement

[a] UCL = upper confidence limit

Figure 1.1 Steps Used in Volumes 1, 2 and 3 to Evaluate Whether a Site has Attained the Cleanup Standard (from Figure 1.1 in Volume 1)



Chapter 2.0 discusses the similarities and differences in Volumes 1, 2 and 3 that are summarized in Table 1.1. Tables 2.1 and 2.2 show the steps in the DQO and DQA processes, respectively. Table 2.3 is a guide for selecting a test from Volumes 1, 2 and 3. Table 2.4 gives a list of assumptions that underlie the statistical tests and procedures in the three volumes. Chapter 3.0 provides discussion and caveats on the use of the volumes. References to the cited literature are given in Chapter 4.0.

There are four appendices to this report. Appendix A provides an introduction to the sampling designs discussed in the volumes. Appendix B presents a detailed list of the assumptions and conditions that underlie the statistical tests described in the volumes. Appendix C provides a case study, Solar Evaporation Basins, to illustrate testing for attainment of a soil background-based standard using methods in Volume 3. Appendix D provides a case study, Process Trenches, to illustrate testing for attainment of a risk-based groundwater standard using methods in Volume 2.

We note that Volumes 1, 2 and 3 do not discuss testing for the attainment of a background-plus-risk (BPR) standard. This type of standard is stated as "X units greater than background," where X is a fixed risk-based standard and "background" is some function or summary statistic of a data set of background measurements, e.g., the arithmetic mean. If the standard is denoted by S, then S = X + background. An example of a BPR standard is on page 602 of USEPA (1983):

> " The concentration of radium-226 in land averaged over any area of 100 square meters shall not exceed the background level by more than 5 pCi/g, averaged over the first 15 cm of soil below the surface... ."

Statistical tests for this type of standard can be performed by adding the constant X to each background measurement and using the tests in Volume 3 to compare these adjusted background data to the site data. In the above example, X is 5 pCi/g.

We also note that tests in Volumes 1 and 2 do not take into account the effects of uncertainty in the risk-based standard (see e.g., NRC (1994) and Gilbert (1994)). Also, the tests in Volume 2 are appropriate only to evaluate the attainment of groundwater risk-based standards, not groundwater reference or background standards. Tests to evaluate attainment of groundwater background standards are discussed by Gibbons (1994) and Davis (1994). Consultation with a statistician familiar with these references is recommended before any statistical test is applied to assess attainment of groundwater background standards. More generally, regardless of the type of cleanup standard or the environmental media of concern, an experienced statistician should be a member of the team that implements the DQO process, the DQA process, and the data analysis and interpretation phases of the study.

CHAPTER 2. SIMILARITIES AND DIFFERENCES IN VOLUMES 1, 2 AND 3

This section discusses the similarities and differences in Volumes 1, 2 and 3 that are summarized in Table 1.1.

2.1 Cleanup Standards

A **risk-based standard** for a contaminant is a specified fixed concentration value that is assumed to be known with certainty. This standard is usually determined at least in part on the basis of human health and/or ecological risk assessment. When testing for the attainment of a risk-based standard, a data set is collected from the remediated site and compared quantitatively in some way to the fixed risk-based standard.

A **reference-based or background-based standard** is based on the *distribution* of the pollutant in the reference or background area. To conduct a test, a data set from the remediated site is compared with a data set from the reference or background area. The tests in Volume 3 are designed to detect when the distribution of measurements for the remediated site is shifted in part or in whole to the right (to higher values) of the reference distribution.

As noted in Chapter 1.0, Volumes 1 and 2 apply to risk-based standards for soil (and solid media) and groundwater, respectively, whereas Volume 3 applies to reference-based or background-based standards for soil and solid media. The type of standard selected must be taken into account when selecting a statistical test to assess attainment of the standard. The tests discussed in Volumes 1 and 2 cannot be used to test for attainment of a reference-based or background-based standard. Similarly, the tests in Volume 3 cannot be used to test for attainment of risk-based standards. Different types of statistical tests are used for risk-based and reference-based standards because only one data set (from the site) is needed for the former, while two data sets (one each from the site and the reference area) are needed for the later.

2.2 Data Quality Objectives (DQO) Process

A common feature of Volumes 1, 2 and 3 is their use of the seven-step DQO process for planning the sampling, measurement and statistical testing program. This process is summarized in Table 2.1 and is more extensively described in USEPA (1993, 1994a). Figure 1.1 indicates where the DQO process fits into the process used to test for attainment of standards. As stated in USEPA (1994a, page 1):

> *The DQO Process* is a strategic planning approach based on the Scientific Method that is used to prepare for a data collection activity. It provides a systematic procedure for defining the criteria that a data collection design should satisfy, including when to

Table 2.1 The Data Quality Objectives (DQO) Process (USEPA 1993, 1994a)

Step 1: State the Problem Concisely describe the problem to be studied. Obtain and review information from prior studies, site history, etc., as needed.
Step 2: Identify the Decision Identify the decisions that must be made and the actions that may result to address the problems identified in Step 1.
Step 3: Identify Inputs to the Decision Identify the information, samples and measurements that are needed to make the decisions identified in Step 2.
Step 4: Define Study Boundaries Specify the time periods, spatial areas, and scales of data aggregation to which the decisions will apply.
Step 5: Develop a Decision rule Define the statistical parameters of interest, specify the action level (standard), and integrate the outputs of DQO Steps 1 through 4 to develop an if-then statement that describes the logical basis for choosing among alternative decisions and actions. Translate that decision rule into hypotheses that can be tested quantitatively.
Step 6: Specify Tolerable Limits on Decision Errors Determine the maximum probabilities of making test decision errors, i.e., specify the decision error rates that can be tolerated by the decision makers. Do this by considering the consequences of making incorrect decisions if the concentrations of the contaminant are at selected values less than, equal to, or greater than the standard.
Step 7: Optimize the Design Evaluate information and outputs from Steps 1 through 6 and generate alternative designs for where and when to collect samples. Choose the most resource-effective design that meets all DQOs. If all DQOs cannot be met within budget, re-examine the DQOs and tolerable decision error limits to consider what tradeoffs are acceptable to the decision makers.

collect samples, where to collect samples, the tolerable level of decision errors for the study, and how many samples to collect.

Volumes 1, 2 and 3 were written before the DQO process was developed to its present level. Hence, the terminology used in the three volumes differs somewhat from that used in USEPA (1993, 1994a). However, the planning process given in the three volumes is still appropriate, and USEPA (1993, 1994a) should be used in conjunction with those volumes.

2.3 Data Quality Assessment (DQA) Process

Once the DQOs are established and samples have been collected and measured, the DQA process (USEPA 1995a, Michael 1993) is used to assess whether the DQOs have actually been attained. The steps of the DQA process are given in Table 2.2. As stated in USEPA (1995a, page 0-1):

> *Data Quality Assessment* is the scientific and statistical evaluation of data to determine if the data are of the right type, quality, and quantity to support their intended use.

The DQA process was developed after Volumes 1, 2, and 3 were written. Nevertheless, the DQA process should be used to assure that the data and measurements obtained do indeed meet the specified DQOs.

2.4 Target Population

The first four steps of the DQO process (Table 2.1) set the stage for decision making. The problems to be addressed have been defined (Step 1), the decisions that must be made to solve the problems have been determined (Step 2), information inputs needed to make decisions have been identified (Step 3), and the geographical and temporal boundaries that define space/time regions for which decisions will be made have been set (Step 4).

In Step 4, the "target population" is defined to ensure that samples are collected for the appropriate space/time domain of the environment. The target population is the geographical and/or time segment of the environment for which a decision will be made about the attainment of standards. Step 4 is important because test results may not apply to geographical areas or time periods that are outside the defined target population. This step is discussed separately here for soil and solid media, and for groundwater.

<u>Soil and Solid Media</u>

> The methods in Volumes 1 and 3 are appropriate when it is feasible to collect soil or solid media samples at random locations (or systematically on a grid) throughout the entire area being evaluated (the target population). The target population used in Volumes 1 and 3 is the set of all possible soil (or solid media) samples from the defined geographical areas being evaluated. It is assumed that concentrations in soil

Table 2.2 The Data Quality Assessment (DQA) Process (USEPA 1995a, Michael 1993)

Step 1: Review DQOs and the Sampling Design After the DQO process is completed, review the outputs of the DQO process to assure that they are still applicable. Review the sampling design and data collection documentation for consistency with the DQOs. If DQOs have not been developed, specify them now and determine if the data collected meet those DQOs.
Step 2: Conduct a Preliminary Data Review Review quality assurance reports, calculate basic statistical quantities (e.g., means and standard deviations) and generate graphs of the data to learn about the structure of the data and to identify patterns, relationships, or potential anomalies.
Step 3: Select the Statistical Test Based on the preliminary data review, check that the most appropriate procedures for summarizing and analyzing the data have been selected. Identify the key underlying assumptions that must hold for the selected statistical tests to be defensible.
Step 4: Verify the Assumptions of the Statistical Test Evaluate whether the underlying assumptions hold, or whether departures are acceptable, given the actual data and other information about the study.
Step 5: Perform the Statistical Test Perform and document the calculations required for the statistical test and document the resulting inferences. If the sampling design may be used again, evaluate the performance of the sampling design and make recommendations regarding its use (and that of competing designs) in the future at this and other sites.

and solid media do not change over time during the sampling phase or after soil remedial action has occurred.

Groundwater

The target population of groundwater that is used in Volume 2 is the set of all possible groundwater samples that could be collected from *existing* wells during the applicable time period. An assumption used in Volume 2 is that existing wells may not adequately represent the entire aquifer. For that reason, the results of tests in Volume 2 should be considered to apply only to wells from which the samples were actually collected, and not to the entire aquifer. Conclusions about the entire aquifer must be based on a combination of the statistical testing results and expert knowledge about hydrogeology, land use, past activities in the area, and other factors that could affect concentrations of contaminants throughout the entire aquifer.

2.5 Decision Errors

The 5th step in the DQO process is for decision makers and the DQO planning team to develop an if-then decision rule for deciding if the site has attained the cleanup standard. The statistician on the planning team will transform this rule into a formal test of hypothesis.

Two types of decision errors are possible when a test of hypothesis is conducted:

- *The test may indicate that the site has attained the standard when it has not*
  - This error may result in a failure to take needed remedial action, which could lead to an increased risk for those who are exposed to contaminants from the site.
- *The test may indicate that the site has not attained the standard when it has*
  - This error may result in taking unnecessary remedial action at the site, thereby expending funds that could be used at sites that really do require remedial action.

The 6th step of the DQO process is for decision makers to define the tolerable limits (probabilities) of making these two decision errors. These limits are necessary to determine the number of sample measurements needed for the test.

The tolerable limits on decision errors should be determined by the decision makers (*not* by the statistician) after considering the consequences of making each type of error when concentrations at the site are assumed (for planning purposes) to be less than, equal to, or greater than the standard. For example, if most concentrations in the target population are greater than the standard, the risk to those exposed to the contamination could be high.

Hence, the decision makers may decide that if in fact most of the concentrations in the target population are larger than the standard, then the probability of failing to detect that the site does not meet the standard should be a rather small value. Of course, decision makers should be aware that if a very small probability is specified, the number of samples required may be large, depending on the variance among the data. USEPA (1993, 1994a) provide further discussion of Step 6.

2.6 Hypotheses Used in Volumes 1 and 2

Once decision makers have defined the tolerable limits on decision errors, the statistician will translate that information into an appropriate null hypothesis (H_0) and alternative hypothesis (H_a) that will be tested by a suitable statistical test. Initially, i.e., before the test is conducted, H_0 is assumed to be true. The statistical test will indicate whether the data is convincing beyond a reasonable doubt that H_0 is incorrect and should be rejected in favor of H_a. Readers who desire a more thorough discussion of hypothesis testing than can be provided here are directed to Kraemer (1985).

Our primary reasons for discussing hypotheses here are that 1) they are set up differently for Volume 3 than for Volumes 1 and 2, and 2) early in the DQO planning process all stakeholders must agree on how H_0 and H_a should be set up for the site being studied. The way in which the hypotheses are set up is determined by which of the two types of decision errors discussed above is most important to avoid. In Volumes 1 and 2, which deal with risk-based standards, the hypotheses used are:

H_0: Cleanup standard has not been attained (1)
H_a: Cleanup standard has been attained

Hence, with this setup, it is initially assumed that the risk-based standard has not been attained and we require strong evidence (beyond a reasonable doubt) that H_0 is incorrect before we conclude with confidence that the cleanup standard has really been attained. That is, the burden of proof is on showing that the standard has been attained. This approach is protective of human health and the environment, i.e., it is deemed more important to guard against unnecessary human or environmental risk than to avoid unnecessary expenditure of funds for unneeded remedial action. More generally, H_0 and H_a in Equation (1) are appropriate whenever the potential consequences (human or ecological health, economic and social costs, etc.) of incorrectly deciding the site has attained the cleanup standard are greater than the potential consequences of incorrectly deciding that the site has not attained the standard.

The use of the hypotheses in Equation (1) in Volumes 1 and 2 means that the tests in those volumes will not have a high probability of indicating that the site has attained the standard unless most concentrations at the site are smaller than the risk-based standard. Hence, some unnecessary cleanup is deemed to be acceptable and worth the environmental and societal costs to guard against the adverse consequences of not detecting when the site has not attained the risk-based standard.

2.7 Hypotheses Used in Volume 3

Volume 3 uses the following hypotheses, which are the reverse of those [Equation (1)] used in Volumes 1 and 2:

$$\begin{aligned} &H_o\text{: Cleanup standard has been attained} \\ &H_a\text{: Cleanup standard has not been attained} \end{aligned} \qquad (2)$$

Therefore, the testing philosophy used in Volume 3 is to initially assume that the site has attained the cleanup standard. Hence, the evidence must be strong (beyond a reasonable doubt) that the cleanup standard has really not been attained before the test will so indicate. That is, the burden of proof is on showing that the standard has not been attained.

The hypotheses in Equation (1) were not used in Volume 3 because their use would result in a large probability that a statistical test would indicate the background standard has not been achieved when in fact concentrations at the site and background are the same. Hence, using the hypotheses in Equation (2) reduces the number of sites where unnecessary additional remedial action is taken. If the hypotheses in Equation (1) were used to test for attainment of background standards, then the site concentrations would have to be considerably less than those at the background site before the test would have a high probability of detecting that fact. That requirement is inconsistent with the reasonable perspective of considering a background-based standard to be an ALARA (As Low As Reasonably Achievable) standard.

A consequence of using Equation (2) rather than Equation (1) for background standards is that sites with concentrations slightly higher than those in the background area will not be identified with high probability as needing cleanup. However, if the appropriate number of samples is used, that problem is not present when site concentrations are substantially higher than those in the background area, which is the case of most concern.

2.8 Sampling Plans

The 7th and last step of the DQO process (Table 1.2) is to use the information developed during the first 6 steps to develop a sampling plan that, with minimal cost, will achieve the decision error limits specified by the decision makers for the test(s) selected. Volumes 1, 2 and 3 provide formulas to determine the number of samples to collect and measure for the selected statistical test. The test decision error rates are inputs to those formulas.

The formulas used to determine the number of samples depend on the particular sampling design selected. Volumes 1 and 2 consider four sampling designs:

- Simple random sampling
- Systematic sampling
- Stratified random sampling
- Sequential sampling

Appendix A provides a brief discussion of these designs and three innovative sampling designs (double sampling, adaptive sampling, and ranked set sampling) that are not discussed in Volumes 1, 2 and 3.

Volumes 1 and 3 also discuss composite sampling, which has the potential for estimating means and totals with less cost and/or greater precision when it is used in combination with any of the above designs. Composite sampling is discussed, e.g., in USEPA (1995b), Edland and van Belle (1994), Lovison, Gore and Patil (1994) and Gilbert (1987).

A systematic triangular grid is the preferred sampling design in Volume 3, although this design, or any grid design, can lead to increased decision errors if the grid points (sampling locations) happen to correspond to a pattern of high or low concentrations. A knowledge of past site operations and remedial actions, or an analysis of previously collected site data, may indicate that this problem is present. To guard against this problem, simple random sampling or an unaligned grid pattern (see Gilbert 1987, p. 93) may be used instead of grid sampling.

2.9 Nondetects

In some cases, measurement methods are not sufficiently sensitive to detect very small concentrations of contaminants with certainty. In those cases, the measurement may be reported by the analytical laboratory as zero or ND (not detected), or as being less than the detection limit for the sample. These "data" are commonly referred to as nondetects or less-than values. Data sets containing nondetects are said to be censored on the left.

Most of the statistical tests in Volumes 1 and 2 require that nondetects be replaced with a reasonable approximate value before the test can be computed. The procedure recommended in Volumes 1 and 2 is to replace each nondetect value with the detection limit for that individual sample. However, other substitutions such as one-half the detection limit might also be considered. In practice, the sensitivity of the statistical test conclusion to the type of substitution should be determined by re-computing the test for each substitution method being considered. If the test conclusion depends on which method is used, then the data or the proposed test may be judged inadequate for making a decision. The authors of Volume 3 do not recommend replacing nondetects with substitute values (such as the detection limit), because the tests described in Volume 3 can be conducted even when a small to moderate number of nondetects is present in one or both data sets. However, certain assumptions must hold even in that situation.

The data analysis of censored data sets is a non-trivial problem that is best handled by a statistician who has experience dealing with this problem. The reader may refer to Volumes 1, 2 and 3 as well as to Gilbert (1995), Helsel (1990), and Gilbert (1987) for further discussion and guidance.

2.10 Statistical Tests

The statistical tests discussed in Volumes 1, 2 and 3 are listed at the bottom of Table 1.1. The test of choice in a given situation will depend on whether:

- the standard to be used is risk-based or background-based
- the attainment of the standard will be evaluated using the estimated mean, median, high percentile of the data distribution, or some other characteristic of the data
- the assumptions that underlie the tests (e.g., uncorrelated and/or normally distributed data) are defensible
- the power of the test (ability to reject H_0 when H_0 is really false) is greater than that of competing tests for a given number of samples
- the performance of the test is appreciably affected by the presence of non-detects in the data set

Table 2.3 is a guide for selecting a test from those discussed in the three volumes. The tests are listed along the top of the table. A dot is used in the table to indicate those tests that can be used when the criteria listed in the left-hand column of the table are fulfilled. We now illustrate the use of Table 2.3.

Example 1

Suppose:

- we need to know if the average concentration of a contaminant in soil at the remediated unit is substantially larger than the average in the background area
- we also need to locate any unsuspected hot spots in the remediated unit
- soil samples will be collected in both the remediated unit and in the background area
- the data are not expected to be normally distributed, or we do not wish to make that assumption
- a large number of nondetects are not expected to occur in either data set

Then it is appropriate to apply more than one statistical test: the Wilcoxon Rank Sum test (to check for changes in average concentrations) and the Quantile test and Hot Measurement comparison (to check for hot spots), as discussed in Volume 3.

Table 2.3 A Guide for Selecting a Test from Volumes 1, 2 and 3.

CRITERIA	UCL on Mean (1,2)*	UCL on Percentile (1,2)	UCL on Proportion (1,2)	No. of Data Exceeding Standard (1)	WRS** Test (3)	Quantile Test (3)	Hot Measurement Test (3)
Applicable to a Risk-Based Standard	•	•	•	•			•
Applicable to a Reference-Based Standard					•	•	•
Detecting Hot Spots		•	•	•		•	•
Detecting Changes in Averages	•				•		
Normal Distribution Required	•	•					
Only Small Number of Non-detects allowed	•	•					
Moderate Number of Nondetects can be Handled			•	•	•	•	•
Test is Easy to Use			•	•		•	•

* Indicates the EPA report volume in which the test is discussed.
** Wilcoxon Rank Sum test

Example 2

Suppose:

- we want a test that will detect when the true mean of well water for a defined time period in a groundwater well exceeds a risk-based standard
- the data are expected to be normally distributed
- no nondetects are expected to occur

Then, it is appropriate to use the data to compute the UCL on the mean concentration for the well and compare it to the groundwater risk-based standard. The procedure that should be used is discussed in Volume 2.

2.11 Assumptions

Many assumptions underlie the statistical tests in Volumes 1, 2 and 3. A list of some key assumptions is given in Table 2.4 for quick reference. A more complete list is provided in Appendix B. Again, a statistical test should be used only when the assumptions that underlie the test are verified during the DQA process as being applicable to the site in question.

Table 2.4 A List of Key Assumptions that Underlie Statistical Tests and Procedures in Volumes 1, 2 and 3.

Assumptions	Vol. 1	Vol. 2	Vol. 3
Cleanup standard is greater than the detection limit	•	•	•
Contamination levels are in a steady state after remediation ceases and before attainment samples have been collected	•	•	•
Trends over time do not exist in the data after attainment samples have been collected	•		•
Type of data distribution is assumed known	•	•	
Spatial correlation is negligible	•	•	•
DQO and DQA processes are used	•	•	•
The remediated unit attains the standard only if all contaminants tested attain their standards	•	•	•
Conclusions of test for attainment of groundwater standards apply only to groundwater wells, not to the entire aquifer		•	
A suitable background area is identified and appropriately sampled			•
Concentrations in the background area are not a significant risk to humans or the environment			•
The same sample collection/measurement procedures are used in the background area and in the remediated site			•
Groundwater measurements follow an autoregressive process model		•	
Location of groundwater wells has been determined by experts in groundwater hydrology		•	
Groundwater measurements taken over time are not correlated after time trends and seasonal cycles have been removed		•	

• Indicates that the assumption is used.

CHAPTER 3. CONCLUDING REMARKS

This document provides an overview of three EPA reports (USEPA 1989, 1992 and 1994b) that provide statistical tests to assess the attainment of risk-based or reference-based (background-based) standards for soil and soil media or groundwater. Two case studies are provided (Appendices C and D) that illustrate some of the testing procedures in the context of the DQO process.

Although the volumes provide tests that will be appropriate in many situations, in practice there may be sites and situations where alternative tests may be more appropriate. Hence, the volumes should not be used as cookbooks that must be followed without question in all cases. To ensure defensible data-driven decisions, the best approach is for the team members, decision makers, and regulators to work together as an effective team to apply the DQO and DQA processes, followed by appropriate data analyses, interpretation of testing results, and thorough documentation.

CHAPTER 4. REFERENCES

Berthouex, P.M. and I. Hau. 1991. "A Simple Rule for Judging Compliance Using Highly Censored Samples," *Research Journal of the Water Pollution Control Federation* 63(6): 880-886.

Conover, W.J. 1980. *Practical Nonparametric Statistics*, 2nd edition, John Wiley & Sons, Inc., New York, NY.

Davis, C.B. 1994. "Environmental Regulatory Statistics," *Handbook of Statistics, Vol. 12, Environmental Statistics*, pp. 817-865, North-Holland, New York.

Edlund, S.D. and G. van Belle. 1994. "Decreased Sampling Costs and Improved Accuracy with Composite Sampling," *Environmental Statistics, Assessment, and Forecasting*, pp. 29-56, Lewis Publishers, Boca Raton.

Gibbons, R.D. 1994. *Statistical Methods for Groundwater Monitoring*, Wiley, New York.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*, Van Nostrand Reinhold, New York, NY.

Gilbert, R.O. 1994. "An Overview of Statistical Issues Related to Environmental Cleanup," In: *Handbook of Statistics, Vol. 12*, (G.P. Patil and C.R. Rao, eds.), pp. 867-880, North Holland Publishers, New York, NY.

Gilbert, R.O. 1995. *A Review of Statistical Methods for Data Sets With Multiple Censoring Points*, PNL-10652, Pacific Northwest National Laboratory, Richland, WA.

Hardin, J.W. and R.O. Gilbert. December 1993. *Comparing Statistical Tests for Detecting Soil Contamination Greater Than Background*, PNL-8989, Pacific Northwest Laboratory, Richland, WA.

Helsel, D.R. 1990. "Less than Obvious," *Environmental Science and Technology* 24(12):1766-1774.

Kraemer, H.C. 1985. "A Strategy to Teach the Concept and Application of Power of Statistical Tests." *Journal of Educational Statistics* 10(3):173-195.

Lovison, G., S.D. Gore and G.P. Patil. 1994. "Design and Analysis of Composite Sampling Procedures: A Review," *Handbook of Statistics, Vol. 12, Environmental Statistics*, pp. 103-166, North-Holland, New York.

Michael, D. 1993. *Data Quality Assessment: Have You Met Your DQOs*? Neptune and Company, Inc. Los Alamos, NM.

National Research Council. 1994. *Science and Judgment in Risk Assessment*, National Academy Press, Washington, D.C.

Patil, G.P., A.K. Sinha and C. Taillie. 1994. "Ranked Set Sampling," *Handbook of Statistics, Vol. 12, Environmental Statistics*, pp. 167-200, North-Holland, New York.

Sprent, P. 1993. *Applied Nonparametric Statistical Methods*, Chapman & Hall, New York.

Thompson, S.K. 1992. *Sampling*, John Wiley & Sons, Inc., New York.

U.S. Environmental Protection Agency (USEPA). 1983. *Standard for Remedial Actions at Inactive Uranium Processing Sites: Final Rule (40 CFR Part 192)*. Federal Register 48(3):590-604.

U.S. Environmental Protection Agency (USEPA). 1989. *Methods for Evaluating the Attainment of Cleanup Standards, Volume 1: Soils and Solid Media*. EPA 230/02-89-042, Office of Policy, Planning, and Evaluation, Washington, D.C.

U.S. Environmental Protection Agency (USEPA). 1991. *A Guide: Methods for Evaluating the Attainment of Cleanup Standards for Soils and Solid Media*, EPA Publication 9355.4-04FS, Office of Emergency and Remedial Response, Washington, D.C.

U.S. Environmental Protection Agency (USEPA). 1992. *Methods for Evaluating the Attainment of Cleanup Standards, Volume 2: Ground Water*. EPA 230-R-92-014, Office of Policy, Planning, and Evaluation, Washington, D.C.

U.S. Environmental Protection Agency (USEPA). 1993. *Data Quality Objectives Process for Superfund, Interim Final Guidance*, EPA/540/G-93/071, Pre-Publication Copy, Office of Emergency and Remedial Response, Washington, D.C.

U.S. Environmental Protection Agency (USEPA). September 1994a. *Guidance for the Data Quality Objectives Process*, EPA QA/G-4, Final, Quality Assurance Management Staff, Washington, D.C.

U.S. Environmental Protection Agency (USEPA). 1994b. *Methods for Evaluating the Attainment of Cleanup Standards, Volume 3: Reference-Based Standards for Soils and Solid Media*, EPA 230-R-94-004, Office of Policy, Planning, and Evaluation, Washington, D.C.

U.S. Environmental Protection Agency (USEPA). 1995a. *Guidance for Data Quality Assessment*, EPA QA/G-9, External Working Draft, Quality Assurance Division, Washington, DC.

U.S. Environmental Protection Agency (USEPA). 1995b. *EPA Observational Economy Series, Volume 1: Composite Sampling*, EPA-230-R-95-005, Office of Policy, Planning, and Evaluation, Washington, DC.

U.S. Environmental Protection Agency (USEPA). 1995c. *EPA Observational Economy Series, Volume 2: Ranked Set Sampling*, EPA-230-R-95-006, Office of Policy, Planning, and Evaluation, Washington, DC.

APPENDIX A

FIELD SAMPLING DESIGNS

This appendix provides brief descriptions of field sampling designs. The following designs are discussed in Volume 1, 2 or 3:

- Simple random sampling
- Systematic sampling
- Stratified random sampling
- Sequential sampling

Simple random sampling consists of selecting sampling points using a random procedure such that every possible sampling point has an equal chance of being selected. Hence, each point is selected independently of the location of other points.

Systematic sampling consists of taking samples in a systematic pattern over the site such as on a square, rectangular, or triangular grid. This design provides uniform coverage and is simple to implement, but the data so obtained can be very misleading if the grid pattern happens to correspond to cyclical or periodic patterns of contamination at the site.

Stratified random sampling consists of dividing the site into relatively homogeneous subareas (strata) and using simple random or systematic sampling in each stratum. This design can result in a better estimate of the overall site mean, which will tend to reduce the probability that a test using the estimated overall site mean will yield an incorrect result for the site.

Sequential sampling consists of conducting the statistical test every time a new individual sample (or samples) becomes available. For the other designs discussed above, the test is conducted only after all the data are available. Volume 1 (page 8-4) suggests that sequential sampling could be implemented by first collecting a group of samples using simple random sampling. Then the samples in the group could be chemically analyzed in random order and the statistical test computed as each new datum becomes available. Sequential sampling permits the termination of field sampling as soon as the test indicates that a decision can be made. The sequential procedure is especially helpful when contamination is very high or very low. Although theoretically there is no upper bound on the total number of samples that may be required to reach a decision, the statistical tests in Volumes 1 and 2 that use sequential sampling are structured to guarantee that a decision will be reached. The expected number of samples required to reach a decision using sequential sampling is usually less than needed when non-sequential sampling is used, except when contamination levels are close to the risk-based cleanup standard. The potential for reducing the number of samples is an important advantage of sequential sampling.

Innovative sampling designs that are not discussed in the three EPA reports should also be considered. The three designs discussed below have potential for reducing costs and/or decision error rates as compared to simple random sampling.

- **Double sampling** (Gilbert 1987), wherein an appropriate number of measurements of two different qualities and costs are obtained to provide improved estimates of means and totals. An example is combining many inexpensive in-situ field measurements with a few expensive but more accurate laboratory measurements to estimate the mean.

- **Adaptive sampling** (Thompson 1992), wherein the procedure for selecting future sampling locations depends on data values obtained previously during the study. An example is selecting additional samples in the geographical neighborhood of each sample whose measured concentration exceeded some upper limit. The original and additional measurements are then mathematically combined so as to obtain an unbiased estimate of the mean concentration.

- **Ranked set sampling** (USEPA 1995c, Patil et al. 1994), which involves first ranking (ordering) randomly selected sampling locations on the basis of expected concentrations at each location. After the locations are ranked, a specific procedure is used to select a few of the ranked locations to be measured. The ranking may be done using expert judgment based on knowledge of operational history at the site, previous data obtained at the site, inexpensive auxiliary measurements, visual inspection of sampling units, or some combination of these methods. Ranked set sampling has been shown to yield better estimates of mean concentrations than simple random sampling. The method has considerable potential for reducing the cost of field sampling efforts associated with testing for the attainment of cleanup standards.

APPENDIX B

ASSUMPTIONS UNDERLYING STATISTICAL PROCEDURES IN VOLUMES 1, 2 AND 3

This appendix provides a list of the assumptions and conditions specified in Volumes 1, 2 and 3 that underlie the tests and testing procedures described in those volumes.

Volumes 1, 2 and 3

- Applicable predictive models and data collected prior to sampling are used in the DQO process to guide development of the sampling design for evaluating attainment of standards.
- Regulatory agencies have specified the contaminants to be evaluated for attainment of standards at the site.
- A suitable quality assurance program for sample collection, handling, and measurement will be used to obtain data to test for attainment of standards.
- Contamination levels are in a steady state (not tending to increase or decrease over time or space) after remediation treatment ceases and before testing for attainment of standards is conducted. Volume 2 (Chapter 7) provides several tests for changes and trends to evaluate this assumption.
- Contaminant concentrations in soil and groundwater populations being sampled do not change after attainment samples have been collected.

  NOTE: Although Volumes 1 and 3 do not discuss tests for trend, such tests are recommended for soil remediation treatments whose performance over time is not well established. After the cleanup standard has been attained, periodic sampling to monitor for unanticipated problems and the validity of assumptions is recommended, particularly for groundwater.
- The site attains the soil cleanup standard only if each contaminant tested attains its standard.
- Consultation with a statistician is recommended if measurements of composite samples will be used to test for attainment.
- The amount of missing data (as opposed to data below the detection limit) is minimized by careful planning, backup procedures, chain of custody procedures, packing, labeling, and record keeping. Provision is made for collecting more samples than required to guard against missing data.

Volumes 1 and 2

- The risk-based standard value for each contaminant in soil or groundwater has been specified by the regulators or stakeholders.
- The cleanup standard is greater than the detection limit.

  NOTE: If the risk-based standard is less than the detection limit then a testing approach like that discussed in Berthouex and Hau (1991) may be considered.

Volumes 1 and 3

- Spatial correlation is small enough to have a negligible impact on test results. Volume 1 (Chapter 10) provides an introduction to geostatistical methods that may be used when significant spatial correlation occurs.
- If composite soil samples are used, all stakeholders have agreed that a measurement obtained from the selected type of composite sample is the appropriate metric for testing the attainment of cleanup standards.

Volume 2

- The location of groundwater wells that are sampled for testing the attainment of standards has been specified by experts in groundwater hydrology and has been approved by regulatory agencies familiar with contamination data at the site.

  NOTE: Because these selected wells are not determined using a probability-based sampling plan, the conclusions from the tests apply only to the water from the selected wells and not to the aquifer in general.

  NOTE: Volume 2 (Chapter 3, page 3-3) states "Extending inferences from the sampled wells to the ground water in general must be made on the basis of both available data and expert knowledge about the groundwater system and not on the basis of statistical sampling theory."

  NOTE: Volume 2 makes no recommendations on the procedure that should be used to identify or place wells for sampling to test for attainment of standards.
- A test for attainment of the groundwater standard for each constituent being tested is conducted for each selected groundwater well.

  NOTE: All selected wells must pass the test for attainment for all constituents being tested.
- Groundwater contamination measurements taken over time are not correlated after long-term time trends and seasonal cycles have been removed from the data.

  NOTE: Volume 2 (Section 5.6) provides statistical procedures to evaluate this assumption.

Volume 2 (continued)

- All statistical methods in Volume 2 for handling serial correlation in ground water measurements assume that ground water measurements obtained over time follow an autoregressive process (Volume 2, Appendix F, Equation F.7).
- Groundwater is said to attain the cleanup standard only if the contaminant concentrations attain the standard for the predictable future.

  NOTE: Volume 2 (Chapter 5) uses "short-term" testing and estimation methods for interim management decisions regarding the current status of a cleanup effort. These methods are not used to assess the attainment of cleanup standards for future points in time. Volume 2 (Chapters 8 and 9) uses "long-term" decision and estimation statistical methods to make decisions about the predictable future.
- A systematic grid or random sampling program is used to obtain data over time to describe groundwater conditions for specified periods of time, such as during remediation.
- Data collected according to a systematic sampling plan are used to assess attainment of groundwater cleanup standards over the long term (Chapters 8 and 9).
- A minimum of four sample collections of groundwater per year is recommended (Chapter 4, page 4-3).
- Systematic sampling is generally preferred over a simple random sample (Chapter 4, page 4-2).
- When the number of samples is determined before sampling begins, three statistical testing methods are provided to assess the attainment of risk-based standards. These three methods are based on computing an upper confidence limit on the mean of groundwater concentrations. One procedure (Box 8.5, page 8-13) assumes the yearly averages are normally distributed. The second procedure (Box 8.12, page 8-18) assumes the log-transformed yearly averages have a normal distribution. The third procedure (Box 8.16, page 8-21) assumes that the data, after being adjusted for seasonal variation, have a normal distribution. Alternatively, a nonparametric test based on the proportion of contaminated samples from one well or an array of wells can be used to test for attainment of risk-based standards (Section 8.5, page 8-25). This nonparametric test requires no assumptions about the distribution of groundwater data.

  •When samples are collected sequentially in time and a new test (of the selected type) is performed each year, three statistical testing methods are provided to assess the attainment of risk-based standards based on computing upper confidence limits on the mean of groundwater concentrations. One procedure (Box 9.4, page 9-9) assumes the yearly averages are normally distributed. The second procedure (Box 9.12, page 9-15) assumes the log-transformed yearly averages have a normal distribution. The third procedure (Box 9.16, page 9-19) assumes that the data after being adjusted for seasonal variation have a normal distribution. Alternatively, a nonparametric test for attainment of risk-based standards can be used based on the proportion of sequentially

Volume 2 (Continued)

selected samples that exceed the cleanup standard (Section 9.4, page 9-22). This nonparametric test requires no assumptions about the distribution of groundwater data.

- When data are skewed, Equation 8.15 in Chapter 8 of Volume 2 (which uses log-transformed yearly averages) is used to compute the upper one-sided confidence limit for the mean to test for attainment of a risk-based groundwater cleanup standard. Note that Equation 8.15 differs from the upper one-sided confidence limit in Gilbert (1987, Equation 13.13, page 170) that applies when data are drawn from a two-parameter lognormal distribution. The procedure in Gilbert (1987) is theoretically correct for the lognormal case, but it will give biased results for other distributions. The procedure in Volume 2 may be applied to any skewed distribution, but the accuracy of its results will vary from case to case.

Volume 3

- A suitable soil background (reference) area has been selected.
- Testing for attainment of a soil background standard takes into account all background measurements collected over the background site.

  NOTE: A single value is <u>not</u> selected from or computed from the set of background data and used as the background standard.
- The background area does not differ from the remediated site in physical, chemical, or biological characteristics that might cause measurements in the background area and remediated site to differ.
- The reference (background) area contains no contamination from the hazardous waste site being evaluated for attainment of background standards.
- Contaminant concentrations in the background area do not present a significant risk to humans or the environment.

  NOTE: That is, there is no significant risk if the site contains concentrations at background levels.
- The same sample collection and measurement procedures are used at the background area and the remediated site.
- The statistical tests do not require that the background and site data follow a normal, lognormal or any other distribution
- The contaminated site need not be remediated to levels less than those in the background area even when the contaminant of interest is present in the background area from anthropogenic (human-made, non-site) sources of pollution, such as from industry or automobiles.
- Contaminant concentrations in the background area and remediated site do not change after samples to assess attainment of standards are collected in those areas.

APPENDIX C

SOIL CASE STUDY: SOLAR EVAPORATION BASINS

This case study illustrates testing for attainment of a soil background-based standard. The 7-step DQO process is used to present the case study. This example is based on an actual contaminated site, although changes have been made to reduce the length and complexity of the example. The test illustrated was not actually conducted and the numerical standard used is fictitious.

Step 1. State the Problem

The Solar Evaporation Basins consist of four sedimentation and flocculation basins that remain from a water treatment facility that has been demolished. Following demolition, the basins served as a solar evaporation facility for liquid chemical wastes. Beginning in 1973, approximately 9,500 kiloliters of a solution were discharged to the basins during the period of waste operations. The solution consisted primarily of sodium nitrate, with trace amounts of other chemicals, predominantly chromium.

The waste deposited in the basins was reduced in volume by evaporation. Basin 1 was used until nitrates were detected in a monitoring well, indicating a possible leak from the basin. Use of Basin 1 was discontinued in 1978. However, Basins 2, 3 and 4 were used with various liners until 1986. Closure activities to remove known contamination in the basins began in 1986.

Step 2: Identify the Decision

The study team and decision makers decided that an important decision to be made is whether the soil in the vicinity of the basins is contaminated to a greater extent than that in a suitable background area. Other decisions, not considered in this example, also need to be addressed, such as whether the contamination presents a risk hazard.

Step 3: Identify Inputs to the Decision

The following inputs were identified.

- Based on historical records, the primary contaminant of concern is chromium. All chromium present is believed to be in valence state VI. Hence, no separate speciation measurements shall be made.
- No suitable background or basin data were available.
- Few if any non-detects are expected.

Step 4: Define the Study Boundaries

- The target population consists of soil in the vadose zone (area of aeration between the earth's surface and the water table) and beneath and immediately adjacent to the basins in a geographically defined zone of specified length, width, and depth. This zone for which a decision will be made shall be called the "study unit."

- As chromium is the primary contaminant of concern, the variable to be tested is chromium in soil.

Step 5: Develop the Decision Rules

- The decision rule selected is:

  If the true median chromium concentration in the study unit is greater than the median chromium concentrations in background soil

  then the background standard has not been attained and the study unit must be remediated.

Using the criteria in Table 2.3, the **Wilcoxon Rank Sum (WRS) test** (Volume 3, Chapter 6) was selected to implement the decision rule.

Step 6: Tolerable Limits on Decision Errors

The two possible decision errors and the maximum decision error rates that could be tolerated by decision makers are:

- Deciding the study unit does not attain the background standard when it does:

  Maximum tolerable decision error rate = 0.05 (5% chance) when the true median chromium concentrations in background and at the study unit are identical.

- Deciding the study unit does attain the background standard when it does not:

  Maximum tolerable decision error rate = 0.10 (10% chance) if the true median chromium concentration in the study unit exceeds the background area median by $\geq$ 20 mg/kg.

That is, the decision makers could tolerate a probability of at most 0.05 (5% chance) of the test incorrectly indicating that the study unit does not attain the background standard. Also, they could tolerate a probability of at most 0.10 (10% chance) of the test failing to detect when the study unit concentration exceeds the median background level by 20 or more mg/kg.

The null and alternative hypotheses selected to be tested were (see Chapter 6 in USEPA 1994a for instructions on how to determine H_o and H_a):

H_0: the median chromium concentration is the same in the study unit as that in the background soil

H_a: the median chromium concentration in the study unit soil is greater than that in the background soil

That is, it shall be initially assumed that the study unit has attained the background standard. In this example, the decision makers have decided that the burden of proof should be on showing that the background standard is not met.

Step 7: Optimize the Design

The procedure described in Section 6.2 in Volume 3 is used to determine the number of samples that should be collected to meet the established limits on decision errors (0.05 and 0.10) when the WRS test is used. Without going into details here, the total number of samples that should be collected (basin plus background area) is determined using Equation (6.3) in Volume 3.

Once data are collected, the DQA process is used to evaluate whether the assumptions that underlie the WRS test and the procedure used to determine the number of samples are valid. If so, then the WRS test statistic (Equation 6.12 in Volume 3) is computed to determine whether to reject H_0. If H_0 is not rejected, then the Quantile test should be conducted. All test results are then documented and decision makers are briefed on the results and their interpretation.

APPENDIX D

GROUNDWATER CASE STUDY: PROCESS TRENCHES

This case study illustrates testing for attainment of a risk-based groundwater standard, using methods in Volume 2. Similar to the case study in Appendix C, this example is based on a real site, but details have been changed to reduce the length and complexity of the example. The 7-step DQO process is used to present the case study.

Step 1: State the Problem

Two process trenches received liquid effluent from nearby uranium fabrication facilities and laboratories for several years. The trenches are approximately 350 meters long, 3.5 meters deep and 3 meters wide (at the bottom of the trench). The effluent contained uranium and was determined (from groundwater sampling) to have migrated to the groundwater beneath the trenches. Consequently, the release of effluent to the trenches was terminated. The problem is to determine if the groundwater uranium concentrations have decreased to a point where they attain the applicable risk-based standard for uranium in groundwater.

Note that this case study does not incorporate any treatment of the groundwater itself (e.g., pump and treat). Instead, the "treatment" is a termination of effluent input to the process trenches at a single point in time.

Step 2: Identify the Decisions

The study team and decision makers identified two primary decisions that must be made to address the problem:

1. Whether post-treatment groundwater concentrations have stabilized sufficiently to permit a test for attainment to be conducted.

2. If so, whether uranium concentrations in post-treatment groundwater extracted from a strategically placed well attain the site-specific risk-based standard.

In this case study, we assume that the post-treatment groundwater concentrations are in a steady state. Chapter 7 of Volume 2 provides graphical methods and statistical tests to evaluate this assumption. We focus on the 2nd decision, i.e., whether the risk-based standard has been attained.

Step 3: Identify Inputs to the Decisions

The following inputs were identified.

- Historical groundwater monitoring data (total uranium concentrations, pCi/l) were collected at weekly intervals from the selected strategically placed well.

  NOTE: These data will be evaluated to determine if they are adequate for assessing the attainment of the uranium risk-based standard for the site.
- A site-specific risk assessment was conducted to develop the groundwater risk-based limit, which is 23 pCi/l of total uranium.

Step 4: Define the Study Boundaries

- The study team and decision makers decided that the target population is extractable well water adjacent to the strategically placed well.

  NOTE: The test results apply only to the extractable water adjacent to the sampled well, not to the entire aquifer.
- An assessment of the standard for the entire aquifer would require examining data from many wells and applying professional judgment with regard to the hydrogeology of the site and the placement of wells.
- The use of a single well in this case study does not constitute an endorsement of using only a single well to evaluate the attainment of groundwater standards.
- Gibbons (1994) and Davis (1994) should be consulted to supplement the procedures in Volume 2 when the goal is to make inferences regarding the entire aquifer on the basis of data from multiple wells and multiple contaminants.

Step 5: Develop the Decision Rules

- The decision rule selected is:

  If the true yearly mean in the strategically placed monitoring well is less than the risk-based standard (23 pCi/l)

  then the risk-based standard has been attained for that well.

Step 6: Tolerable Limits on Decision Errors

The two possible decision errors and their tolerable decision error limits (determined by the decision makers, *not* the statistician, on the basis of the consequences of decision errors) are:

- Deciding the groundwater in the well does not meet the standard when it does

  Maximum tolerable decision error rate = 0.20 (20% chance) if the true mean uranium concentration in the groundwater is 15 pCi/l or less (15 pCi/l was selected by the decision makers).

- Deciding the groundwater in the well does meet the standard when it does not

  Maximum tolerable decision error rate = 0.05 (5% chance) if the true mean uranium concentration in the groundwater is $\geq$ 23 pCi/l (the standard).

The null and alternative hypotheses selected to be tested were (Chapter 6 in USEPA 1994a gives instructions for how to determine H_o and H_a):

H_o: The groundwater standard is not attained

H_a: The groundwater standard is attained

That is, it shall be initially assumed that the groundwater standard is not attained. Thus, the burden of proof is on showing that the standard has been attained.

Using the criteria in Table 2.3 and the fact that H_o is stated as above, the upper confidence limit (UCL) on the mean of the yearly means was selected to implement the decision rule. Moreover, the 95% UCL must be used, because the tolerable decision error limit was set to a 5% chance of deciding that the groundwater in the strategically placed well does meet the standard when it does not.

Step 7: Optimize the Design

The procedures described in Section 8.2.1 of Volume 2 may be used to determine the number of samples needed for the 95% UCL test to meet the established tolerable limits on the two types of decision errors for the single well.

NOTE: In actual practice, Gibbons (1994) and Davis (1994) should be used to supplement Volume 2. For example, Gibbons (1994) shows how to determine the required number of samples when more than one well and/or contaminant will be tested for attainment of standards.

The required number of samples is compared with the number available for the well to see if more data are needed. This evaluation of past data is part of the DQA process that is conducted to assure that data meet the established DQOs before the data are used to make decisions.

Once application of the DQA process has shown that the data are adequate, the upper one-sided 95% confidence interval on the well mean is computed as shown in Box 8.9 (Section 8.3) of Volume 2. The test result is then documented and decision makers are briefed on the results and their interpretation.