EPA 540-R-01-003
OSWER 9285.7-41
June 2001

Guidance for Characterizing Background Chemicals In Soil at Superfund Sites

External Review Draft

Office of Emergency and Remedial Response
U.S. Environmental Protection Agency
Washington, DC 20460

PREFACE

This document provides guidance to the U.S. Environmental Protection Agency Regions concerning how the Agency intends to exercise its discretion in implementing one aspect of the CERCLA remedy selection process. The guidance is designed to implement national policy on these issues.

Some of the statutory provisions described in this document contain legally binding requirements. However, this document does not substitute for those provisions or regulations, nor is it a regulation itself. Thus, it cannot impose legally binding requirements on EPA, States, or the regulated community, and may not apply to a particular situation based upon the circumstances. Any decisions regarding a particular remedy selection decision will be made based on the statute and regulations, and EPA decision makers retain the discretion to adopt approaches on a case-by-case basis that differ from this guidance where appropriate. EPA may change this guidance in the future.

ACKNOWLEDGMENTS

The EPA working group, chaired by Jayne Michaud (Office of Emergency and Remedial Response), included Tom Bloomfield (Region 9), Clarence Callahan (Region 9), Sherri Clark (past working group chairperson; Office of Emergency and Remedial Response), Steve Ells (Office of Emergency and Remedial Response), Audrey Galizia (Region 2), Cynthia Hanna (Region 1), Jennifer Hubbard (Region 3), Dawn Ioven (Region 3), Julius Nwosu (Region 10), Sophia Serda (Region 9), Ted Simon (Region 4), and Paul White (Office of Research and Development). David Bennett (Office of Emergency and Remedial Response), was the senior advisor for this working group. Comments and suggestions provided by Agency staff are gratefully acknowledged.

Technical and editorial assistance by Drs. Mary Deardorff and N. Jay Bassin of Environmental Management Support, Inc., and Dr. Harry Chmelynski of Sanford Cohen & Associates, Inc., are gratefully acknowledged.

This report was prepared for the Office of Emergency and Remedial Response, United States Environmental Protection Agency. It was edited and revised by Environmental Management Support, Inc., of Silver Spring, Maryland, under contract 68-W6-0046, work assignment 007, managed by Jayne Michaud. Mention of trade names or specific applications does not imply endorsement or acceptance by EPA. For further information, contact Jayne Michaud, U.S. EPA, Office of Emergency and Remedial Response, Mail Code 5202G, 1200 Pennsylvania Avenue, Washington, DC 20460.

CONTENTS

EXHIBITS

CHAPTER 1

INTRODUCTION

The U.S. Environmental Protection Agency (EPA) developed this document to assist Superfund remedial project managers (RPMs) and risk assessors when implementing Superfund baseline risk assessments and the remedy selection process. This document recommends statistical methods for characterizing background concentrations of chemicals in soil for the purpose of evaluating risks and making remedial decisions.

This document supplements Agency guidance included in the *Risk Assessment Guidance for Superfund Vol. I, Human Health Evaluation Manual (Part A)*[1] (RAGS). RAGS contains useful guidance on background issues that the reader should also consult:

- Sampling needs (Sections 4.4 and 4.6)
- Statistical methods (Section 4.4)
- Exposure assessment (Section 6.5), and,
- Risk characterization (Section 8.6).

This document draws upon many other publications and statistical references, which are cited in Chapters 2 through 5. In general, background may play a role in the Superfund process when:

- Determining whether a release falls within the limitation contained in Section 104(a)(3)(A) of the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), which addresses naturally occurring substances in their unaltered form from a location where they are naturally found[2];

- Developing remedial goals[3]; and

- Communicating cumulative risks associated with the Superfund site.

As stated in RAGS, a statistically significant difference between background samples and site-related contamination should not, by itself, trigger a cleanup action. Risk assessment methods should be applied to ascertain the significance of the chemical concentrations. A national policy is being developed to clarify the role of background characterization results in the Superfund risk assessment and remedy selection process. When completed, the policy will be included in this guidance.

1.1 Application of Guidance

Not every Superfund site investigation will need to characterize background chemicals. A background evaluation usually is considered when certain contaminants that pose risks are believed to be attributable to background. The need for background characterization and the required level of effort should be determined on a site-specific basis. The site team should consider whether collecting background samples is necessary (Chapter 2); when, where, and how to collect background samples (Chapter 3); and how to evaluate the data (Chapters 4 and 5).

1.2 Goals

The general goals of this guidance are to:

- Provide a practical guide for characterizing background concentrations at Superfund sites; and

- Present sound options for evaluating background data sets in comparison to site contamination data.

1.3 Scope of Guidance

This guidance pertains to the evaluation of chemical contamination in soil at Superfund sites. This guidance may be updated in the future to address non-soil media. Non-soil media are dynamic and influenced by upstream or upgradient sources. Such media—air, groundwater, surface water, and sediments—typically require additional analyses of release and transport, involve more complex spatial and temporal sampling strategies, and require different ways of combining and analyzing data.[4]

Because this guidance pertains to background chemicals, the user should consult the available Agency guidances and policies when dealing with sites with radioactive contaminants. Certain types of Superfund sites, such as mining or dioxin-contaminated sites, may require consideration of specific Agency policies and regulations. Therefore, this guidance should be applied on a case-by-case basis, with consideration of Agency statutes, regulations, and policies.

1.4 Intended Audience

The intended audience of this guidance is Superfund staff which includes risk assessors, RPMs, and decision makers. To the extent practicable, this guidance may also be applicable to sites addressed under removal actions, especially non-time-critical removal actions, and Resource Conservation and Recovery Act (RCRA) corrective actions.

1.5 Definition of Background

For the purposes of this guidance, background samples are those collected at or near the hazardous waste site in areas not influenced by the Superfund site contamination or other nearby Superfund sites. Background soil samples should have the same basic characteristics as the site sample.

Background substances may be natural or man-made. An example of a naturally occurring substance is arsenic present in soil as a result of natural geologic processes. In some geographic areas, naturally occurring substances may be ubiquitous. Man-made or anthropogenic substances are present in the environment because of human activities. In some geographic areas, man-made substances may be ubiquitous in soil, such as dioxin and pesticides. Some constituents in background soil samples could exist as a result of both natural and man-made conditions (such as naturally occurring arsenic and arsenic from pesticide applications or smelting operations).

Superfund site activity (such as waste disposal practices) may cause naturally occurring substances to be released into other environmental media or chemically transformed. The concentrations of the released naturally occurring substance may not be considered as representative of natural background according to CERCLA 104(a)(3)(A).

CHAPTER NOTES

1. U.S. Environmental Protection Agency (EPA). 1989. *Risk Assessment Guidance for Superfund Vol. I, Human Health Evaluation Manual (Part A).* Office of Emergency and Remedial Response, Washington, DC. EPA 540-1-89-002. Hereafter referred to as "RAGS."

2. CERCLA 104(a)(3)(A) restricts the authority to take an action in response to the release or threat of release of a "naturally occurring substance in its unaltered form or altered solely through naturally occurring processes or phenomena, from a location where it is naturally found."

3. The National Oil and Hazardous Substances Pollution Contingency Plan (NCP) (40 CFR Part 300) is the primary regulation that implements CERCLA. The preamble to the NCP discusses the use of background data for setting cleanup levels for constituents at Superfund sites.

   "...In some cases, background levels are not necessarily protective of human health, such as in urban or industrial areas; in other cases, cleaning up to background levels may not be necessary to achieve protection of human health because the background level for a particular contaminant may be close to zero, as in pristine areas" (55 FR 8717-8718).

The preamble to the NCP also identifies background as a technical factor to consider when determining an appropriate remedial level:

> "Preliminary remediation goals...may be revised to a different risk level within the acceptable risk range based on the consideration of appropriate factors including, but not limited to: exposure factors, uncertainty factors, and technical factors...Technical factors may include...background levels of contaminants..."(55 FR 8717).

4. RAGS Sections 4.5 and 6.5.

CHAPTER 2

DETERMINING THE NEED FOR BACKGROUND SAMPLING DATA

A first step in determining the need for background sampling data is gathering and evaluating all of the available data. Some information gathered during the Preliminary Assessment/Site Investigation (PA/SI) may provide data on suspected background chemicals. The SI usually provides the first opportunity to collect some background samples. Data collected and assessed for the hazard ranking system (HRS) process may include both site-related contaminants and off-site (or estimated background) substances. These data are generally limited in quantity and sample location. The locations of all data need to be identified and reported when these data are considered during the remedial investigation. The general types of information to consider when determining the need for background sampling are listed in Exhibit 2.1.

Information from preliminary site studies or published sources (regional or local data from the state or U.S. Geological Survey) may be useful for identifying local soil, water, and air quality characteristics.[1] Data from these resources may be useful for qualitative analyses of regional conditions. However, usually they are not sufficient to assess site-specific conditions in a quantitative manner. The EPA site team should determine the utility of these data.

After compiling and considering the relevant information, the site team should determine if the data are sufficient for the risk assessment and risk management decisions, or if additional site-specific data should be collected to characterize background.

Background Sampling Considerations

- Natural variability of metals in soil
- Operational practices
- Waste type
- Contaminant mobility
- Soil type(s)

2.1 When Background Samples Are Not Needed

If the sample quantity, location, and quality of

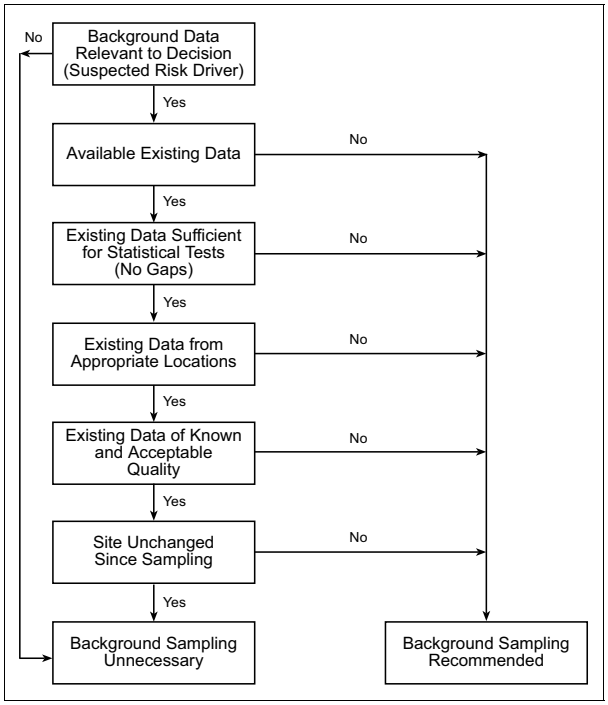

Exhibit 2.1 Determining the need for background sampling.

existing data can be used to characterize background concentrations and compare them to site data, then the team may not need additional samples. In some cases, the team might determine that background concentration levels are irrelevant to the decision-making process. For example, if the team is addressing a chemical release whose constituents are known and not naturally occurring, background data would not be relevant. In other cases, suspected background constituents may not exceed risk-based cleanup goals, and, therefore, further background analysis would not be relevant.

2.2 When Background Samples Are Needed

In some cases, the existing data may be inadequate to characterize background. The reasons for this include, but are not limited to, the following:

- Insufficient number of samples to perform the desired statistical analysis or to perform the tests with the desired level of statistical power;
- Inappropriate background sample locations (such as those affected by another contamination source, or in soil types that do not reflect on-site soil types of interest);
- Unknown or suspect data quality;
- Alterations in the land since the samples were collected (such as by filling, excavation, or introduction of new anthropogenic sources); and
- Gaps in the available data (certain chemicals were excluded from the sample analyses, or certain soil types were not collected).

CHAPTER NOTE

1. U.S. Environmental Protection Agency (EPA). October 1988. *Guidance for Conducting Remedial Investigations and Feasibility Studies Under CERCLA; Interim Final*. (NTIS PB89-184626, EPA 540-G-89-004, EPA 9355.3-01).

CHAPTER 3

DEVELOP THE SAMPLING AND ANALYSIS PLAN

Data Quality Objectives (DQOs) should be used when developing sampling and analysis plans (SAPs) to ensure that reliable data are acquired. RAGS (Chapter 4) describes the role of the DQO process in general terms. The process is outlined here for purposes of developing background sampling plans. For further details, consult Section 6 of *Guidance for the Data Quality Objectives Process*[1] and *Guidance for Data Quality Assessment: Practical Methods for Data Analysis*.[2]

The DQO process is the starting point for many decisions that shape the sampling plan. It involves a series of steps for making optimal decisions based on limited data. A careful statement of the DQOs for a study will clarify the study objectives, define the most appropriate type of data to collect, determine the most appropriate conditions for collecting the data, and specify limits on decision errors. Use of the DQO process ensures that the type, quantity, and quality of environmental data used in decision making will be appropriate for the intended application. It improves efficiency by eliminating unnecessary, duplicative, or overly precise data. The DQO process provides a systematic process for defining a tolerable level for decision errors. The DQO process and decision parameters establish the quantity and quality of data needed.

3.1 DQO Steps for Characterizing Background

Each of the seven steps of the DQO process answers a question phrased in terms of background issues. The examples below should be modified to fit the site of concern. A statistician should be consulted as needed.

Seven Steps in the Data Quality Objectives Process

1. State the Problem
2. Identify the Decision
3. Identify Inputs to the Decision
4. Define Boundaries of Study
5. Develop a Decision Rule
6. Specify Limits on Decision Errors
7. Optimize the Design for Obtaining Data

Step 1. State the Problem: *Example: Are there differences between the concentrations of a contaminant (risk driver) that are found on site and those concentrations that are found off-site (background)?*

Tasks include:

- Identifying the resources available to resolve the problem, including scoping team; and
- Developing or refining the comprehensive conceptual site model.

Step 2. Identify the Decision: *Example: Are the chemical(s) associated with a site-related source or background?*

Tasks include:

- Identifying the chemicals to analyze; and
- Determining if these chemicals are expected to occur in reference areas selected to reflect background conditions.

Step 3. Identify Inputs into the Decision: *Example: What kinds of data are needed? What kinds of data are available?*

Definitions

Δ (delta): The true difference between the concentration of chemical X in contaminated areas and the background concentration of chemical X. Delta is an unknown parameter which describes the true state of nature. Hypotheses about its value are evaluated using statistical hypothesis tests.

S (substantial difference): The largest difference Δ that is acceptable based on risk assessment. S is the action level. If Δ exceeds S, the site requires further evaluation and possible remediation. S is *not* related to the number of samples or their values.

MDD (minimum detectable difference): The smallest difference which the data and statistical test can resolve. The MDD depends on sample-to-sample variability, the number of samples, and the power of the statistical test. The MDD is a property of the survey design.

Gray Region: A range of values of Δ where the statistical test will yield inconclusive results. The width of the gray region is equal to the MDD for the test. The location of the gray region depends on the type of statistical test selected.

Tasks include identifying:

- Which chemicals need to be analyzed;
- Which soil types and depths need to be sampled;
- Which comparison tests are likely to be used (see Chapter 5 for details about comparison tests);
- What coefficient of variation is expected for the data (based on previous samples if possible);
- What preliminary remediation goals (PRGs) or applicable or relevant and appropriate requirements (ARARs) may need to be met; and
- What are the desired power and confidence levels?

Decision outputs for background characterizations are discussed in detail in Chapter 5.

Step 4. Define Boundaries of the Study: *Example: What are the spatial and temporal aspects of the environmental media that the data must represent to support the decision?*

Tasks include:

- Defining the geographic areas for field investigation;
- Defining the characteristics of the soil data or population of interest;
- Dividing the soil data population of interest into strata having relatively homogeneous characteristics;
- Determining the timeframe to which the decision applies; and
- Identifying practical constraints that may hinder sample collection.

Step 5. Develop a Decision Rule: *Example: If the mean concentration in contaminated areas exceeds the mean background concentration, then the chemical will be treated as site-related. Otherwise, if the mean concentration in contaminated areas does not exceed the background mean, the chemical will be treated as coming from the same population as background.*

Tasks include:

- Choosing the null hypothesis, H_0;
- Specifying the alternative hypothesis, H_A;
- Specifying the gray region for the hypothesis test; and
- Determining the level of a substantial difference above background, S.

Hypothesis testing is an approach that helps the decision maker through the analysis of data. Chapter

5 discusses the application of hypothesis testing at Superfund sites. General information on hypothesis testing is provided in Section 3.2.

Step 6. Specify the Limits on Decision Errors: *Example: What level of uncertainty is acceptable for this decision? (For definitions, see Sections 3.2 on Hypothesis Testing and 3.3 on Errors and Confidence, and Exhibits 3.1 and 3.2.):*

- *Test form 1—The gray region extends from a difference of Δ = 0 on the left to Δ = MDD on the right (see box for definitions). Acceptable limits on decision errors are α at the left edge of the gray region, and β at the right edge.*
- *Test form 2—The gray region extends from a difference of Δ = (S - MDD) on the left to Δ = S on the right. The acceptable limits on decision errors are α at the right edge of the gray region, and β at the left edge.*

Tasks include:

- Determining the possible range of the parameter of interest (Δ);
- Specifying both types of decision errors (Type I and Type II—see Section 3.2);
- Identifying the potential consequences of each type of error, specifying a range of possible values for Δ (the gray area—see Exhibits 3.1 and 3.2) where consequences of decision errors are relatively minor; and
- Selecting the limits on decision errors (α and β) to reflect the decision-maker's concern about the relative consequences for each type of decision error (Section 3.3).

Step 7. Optimize the Sampling Design: *Example: What is the most resource-effective sampling and analysis design for generating data that are expected to satisfy the DQOs?*
Tasks include:

- Reviewing the DQO outputs and existing environmental data;
- Developing general sampling and analysis design alternatives;
- Verifying that DQOs are satisfied for each design alternative;
- Selecting the most resource-effective design that satisfies all of the DQOs; and
- Documenting the operational details and theoretical assumptions of the selected design in the sampling and analysis plan.

More information may be required to make a decision. If the required sample size is too large, it may be necessary to modify the original DQO parameters. To reduce survey cost while maximizing utility of the available resources, one or more of the constraints used to develop the survey design may be relaxed. A discussion on adjusting sample size is provided in Section 3.4.

3.2 Hypothesis Testing

The decision rule (DQO Step 5) involves developing a logical "if...then..." statement that defines the conditions that would choose among alternative actions. The first step in developing the decision rule is to transform the problem into statistical terminology by developing a *null hypothesis* and an *alternative hypothesis*. These hypotheses form the two alternative decisions in the decision rule.

Action Levels and Background

In comparisons with background, the parameter of interest is symbolized by the Greek letter *delta* (Δ), the amount by which the distribution of concentrations in contaminated areas exceeds the background distribution. As indicated in the "definitions" box on the next page, Δ is an unknown parameter that represents the true state of nature. Although it is impossible to know Δ exactly, statistical tests are used to evaluate hypotheses made concerning the possible values of Δ. The statistical tests are used to reject or not reject hypotheses about Δ based on test statistics computed from the sample data.

The decision rule requires a parameter of interest (such as mean, median, maximum) and specification of an action level for the decision.

Null and Alternative Hypotheses

In statistics, as in science, a "hypothesis" is a presumption of fact (true or false) that can be tested. Conventionally, hypotheses are stated in such a way that we know what to expect if they are true. For this manual (as in RAGS), this is referred to as the "research hypothesis." However, in order to "prove" the desired hypothesis, it is commonly easier to try to *disprove* it (that is, the hypothesis is *not* true). This assumption to be tested is called the *null hypothesis* (H_0)—if the null hypothesis is true, then the initial presumption is *not true*. If we want to show that site concentration exceeds background, we formulate a null hypothesis that their concentrations are the same. A null hypothesis, then, is any testable presumption set up to be disproved.

An *alternative hypothesis* (H_A) is the logical opposite of the null hypothesis: if H_0 is true, H_A is false, and vice-versa. Consequently, the alternative hypothesis is usually logically the same as the investigator's research hypothesis. However, null and alternative hypotheses may need to be formulated to consider both tails of the curve if it matters whether the statistic of interest is greater than or less than the true mean, not just different from the true mean. Since H_A is the conclusion you draw if you have sufficient evidence to reject H_0, it is usually written as an inequality (*e.g.*, $\mu_s > \mu_b$; $\mu_s < \mu_b$; $\mu_s \neq \mu_b$).

For background comparisons, the parameter of interest is Δ, the difference between the mean concentration in contaminated areas and the mean concentration in background areas. The action level for background comparisons is the largest value of the difference in means that is acceptable to the decision maker. In this guidance, the action level for the difference in means is defined as a substantial difference (S), which may be zero or a positive value based on risk assessment, applicable regulation, or guidance. In some cases, the largest acceptable value for the difference in means may be $S = 0$.

Estimates of Δ are obtained by measuring contaminant concentrations in contaminated areas and in background areas. For example, one estimate of the mean concentration in contaminated areas is the simple arithmetic average of the measurements from these areas. An estimate of the mean background concentration is similarly calculated. An estimate of the difference in means (Δ) is obtained by subtracting the mean background concentration from mean concentration in contaminated areas. In most cases of interest, the estimate of Δ will be a positive number. If there is little or no contamination on the site, then the estimate for Δ may be near zero or slightly negative. Note that the estimated value for Δ calculated by this simple procedure (or by any more complicated procedure) is only an approximation of the true value of Δ. Hence, decisions based on any estimated value for Δ may be incorrect.

Using the DQO process, the decision maker must choose between two courses of action, one associated with the null hypothesis and one associated with the alternative hypothesis. Adopting the DQO approach and hypothesis tests can control the probability of making decision errors. Hypothesis testing is a quantitative method to determine whether a specific statement concerning Δ (called the null hypothesis) can be rejected by examining the data, or not. Decisions concerning the true value of Δ reduce to a choice between "yes" or "no." When viewed in this way, two types of incorrect decisions, or decision errors, may occur:

- Incorrectly deciding that the answer is "yes" when the true answer is "no;" and
- Incorrectly deciding the answer is "no" when the true answer is "yes."

While the possibility of decision errors can never be totally eliminated, it can be controlled. To control decision errors, it is necessary to control the uncertainty in the estimate of Δ. Uncertainty arises from three sources:

- Sampling error;
- Measurement error; and
- Natural variability.

The decision maker has some control of the first two sources of uncertainty. Sampling error may be controlled by collecting a larger number of samples. Larger samples lead to fewer decision errors. Use of more precise measurement techniques or duplicate measurements can reduce measurement error, thus reducing the likelihood of a decision error. However, the third source of uncertainty is more difficult to control. Natural variability arises from the uneven distribution of contamination on the site and in background areas. Natural variability is measured by the true standard deviation (σ) of the distribution of contamination. A larger value of σ indicates that a larger number of measurements will be needed to achieve a desired limit on decision errors. It is important that overly optimistic estimates for σ be avoided because this may result in a design that fails to generate data with sufficient power for the decision.

The DQO process provides a formal procedure to quantify the decision maker's acceptable limits for decision errors. The decision maker's limits on decision errors are used to establish performance goals for data collection. The goal of the DQO process is to develop a data collection plan that reduces the chance of making decision errors of both types. The first step in the DQO process includes specifying the gray region for the test. The gray region is a range of possible values of Δ, where the consequences of making a decision error are relatively minor.

Any useful statistical test has a low probability of reflecting a substantial difference when the site and background distributions are identical (false positive) but has a high probability of reflecting a substantial difference when the distribution of contamination in contaminated areas greatly exceeds the background distribution. In the gray region between these two extremes, the statistical test has relatively poor performance. When the test procedure is applied to a site with a true concentration distribution in the gray region, the test may indicate that the site exceeds background, or may indicate that the site does not exceed background, depending on random fluctuations in the sample data.

It is necessary to specify a gray region for the test because the decision may be "too close to call" due to uncertainty in the estimate of Δ. The second step in the DQO procedure for specifying limits on decision errors is to assign upper bounds on the decision error rates for values of Δ above and below the gray region. These bounds limit the probability of occurrence of decision errors.

The exact definition of the gray region is determined by the type of hypothesis test that is selected by the decision maker (See Exhibits 3.1 and 3.2 in Section 3.4). In general, the gray region for Δ is to the right of the origin ($\Delta = 0$) and bounded from above by the substantial difference ($\Delta = S$). Additional guidance on specifying a gray region for the test is available in *Guidance for the Data Quality Objectives Process*.[2] The expected outputs of the DQO process are the gray region and the decision error limits based on the consequences of making an incorrect decision.

The width of the gray region is called the "minimum detectable difference" for the statistical test, indicating that differences smaller than the MDD cannot be detected reliably by the test. If the test is used to determine if concentrations in the contaminated area exceed background concentrations by more than S, it is necessary to ensure that MDD for the test is less than S. In the planning stage, this requirement is met by designing a sampling plan with sufficient power to detect differences as small as S. If the data were collected without the benefit of a sampling plan, retrospective calculation of the power of the test may be necessary before using the data to make a decision.

In the planning stage, the absolute size of the MDD is of less importance than the ratio of the MDD to the natural variability of the contaminant concentrations in the contaminated area. This ratio is termed the "relative difference" and defined as MDD/σ,

where σ is the standard deviation of the distribution of contamination in contaminated areas. The relative difference expresses the power of resolution of the statistical test in units of uncertainty. Relative differences of less than one standard deviation (MDD/σ < 1) are difficult to resolve unless a large number of measurements are available. Relative differences of more than three standard deviations (MDD/σ > 3) are easier to resolve. The goal for the data collection plan should be to achieve values of MDD/σ between one and three. The required number of samples increases dramatically when MDD/σ is smaller than one. Conversely, little advantage is gained by making MDD/σ larger than three. If MDD/σ is greater than three, additional measurement precision is available at minimal cost by making the width of the gray region (MDD) smaller.

The number of measurements required to achieve the specified decision error rates has a strong inverse relationship with the value of MDD/σ. The cost of the data collection plan should be examined quantitatively for a range of possible values of the MDD before selecting a final value. A tradeoff exists between cost (number of samples required) and benefit (better power of resolution of the test). The tradeoff analysis should begin with analysis of the choice MDD = S, where S is a substantial difference. If the relative substantial difference (S/σ) exceeds three, then a reasonably small number of samples are required for this minimally acceptable test design. Additional measurement precision is available at minimal cost by choosing MDD < S. A binary search procedure would indicate the choice of MDD = S/2 as the next trial in the cost tradeoff comparison. If S/σ is between one and three, then selecting MDD = S is a reasonable alternative. If S/σ < 1, then selecting MDD = S is the most cost-effective choice consistent with the requirement that MDD ≤ S.

The MDD, in conjunction with the values selected for tolerable decision error rates, determines the cost of the survey design produced by the DQO process and the success of the survey in determining which areas present unacceptable risks. From a risk assessment perspective, selection of the proper width of the gray region is one of the most difficult tasks in the DQO process. One goal of the DQO process is to make the MDD as small as possible within the goals and resources of the cleanup effort.

Two forms of the statistical hypothesis test are useful for comparisons with background. The null hypothesis in the first form of the test is that the site is *indistinguishable from background*. The null hypothesis in the second form of the test is that the site *exceeds background by a substantial difference*. RAGS[3] provides guidance for the first form of the background hypothesis test. Both forms are described in the next section.

3.2.1 Background Test Form 1

The null hypothesis for background comparisons, "the concentration in contaminated areas does not exceed background concentration," is formulated for the express purpose of being rejected:

- *The null hypothesis (H_0).* The mean contaminant concentration in samples from contaminated areas is less than or equal to the mean concentration of the background data ($\Delta \leq 0$).[4]

- *The alternative hypothesis (H_A).* The mean contaminant concentration in samples from contaminated areas is significantly higher than the mean of the background data ($\Delta > 0$).

When using this form of hypothesis test, the data must provide statistically significant evidence that the null hypothesis is false—the site does exceed background. Otherwise, the null hypothesis cannot be rejected based on the available data, and the concentrations found in contaminated areas are considered equivalent to background.

An easy way to think about the decision errors that may occur using Background Test Form 1 is to think about the criminal justice system in this country and consider what a jury must weigh to determine guilt. The only choices are "guilty" and "not guilty." A person on trial is presumed "innocent until proven

guilty." When the evidence (data) is clearly not consistent with the presumption of innocence, a jury reaches a "guilty" verdict. Otherwise the verdict of "not guilty" is rendered when the evidence is not sufficient to reject the presumption of innocence. A jury does not have to be convinced that the defendant is innocent to reach a verdict of "not guilty." Similarly, when using Background Test Form 1, the null hypothesis is presumed true until proven false.

Two serious problems arise when using Background Tests Form 1. One type of problem arises when there is a very large amount of data. In this case, the MDD for the test will be very small, and the test will almost always reject the null hypothesis. Even very small differences between the site and background mean concentrations can be resolved in this case. If the site exceeds background by more than an infinitesimal amount, there is a 100 percent chance that the null hypothesis will be rejected if a sufficiently large number of samples is taken. Therefore, the sample size should exceed the minimum number of samples required to give the test sufficient power.

A second type of problem may arise in the use of Background Test Form 1 when insufficient data are available. This may occur, for example, when the on-site or background variability was underestimated in the design phase. In this case, the statistical test is unlikely to reject the null hypothesis due to the lack of sufficient power. When using Background Test Form 1, it is always best to conduct a retrospective power analysis to ensure that the power of the test was adequate to detect contaminated areas that exceed background by more than the MDD. A simple way to do this is to recompute the required sample size using the sample variance in place of the estimated variance that was used to determine the required sample size in the planning phase. If the actual sample size is greater than this post-calculated size, then it is likely that the test has adequate power.[5] If the retrospective analysis indicates that adequate power was not obtained, it may be necessary to collect more samples. Hence, if large uncertainties exist concerning the variability of the contaminant concentration in contaminated areas, Background Test Form 1 may lead to inconclusive results.

Detailed information on the application and characteristics of Background Test Form 1 is available in the document series *Statistical Methods for Evaluating the Attainment of Cleanup Standards.* Volume 3, subtitled *Reference-Based Standards for Soils and Solid Media*[6] contains detailed procedures for comparing site measurements with background reference area data using parametric and nonparametric tests based on Background Test Form 1.

3.2.2 Background Test Form 2

An alternative form of hypotheses test for comparing two distributions is presented in *Guidance for the Data Quality Objectives Process, EPA QA/G-4*.[1] When adapted to the background problem, the null hypothesis, "the concentration in contaminated areas does exceed background concentration," again is formulated for the express purpose of being rejected:

- *The null hypothesis* (H_0): The mean contaminant concentration in contaminated areas exceeds background by more than a substantial difference S ($\Delta > S$).[7]

- *The alternative hypothesis* (H_A): The mean contaminant concentration in contaminated areas does not exceed background by more than a substantial difference S ($\Delta \leq S$).

Here, the *substantial difference* is an action level that reflects a substantial and undesired increase in risk over background risks. Although there is no explicit use of the quantity S in the statement of the hypotheses used in Background Test Form 1, an estimate of S is important for determining an upper limit for the MDD for Background Test Form 2, as discussed below. Issues affecting the determination of site-specific values for a substantial difference are discussed in more detail in the Appendix at the end of this guidance.

Detailed information on the application and charac-

Hypothesis Testing: Type I and Type II Errors

Decision Based on Sample Data	Actual Site Condition	
	H_0 is True	**H_0 is not True**
H_0 is not rejected	Correct Decision: (1 - α)	Type II Error: False Negative (β)
H_0 is rejected	Type I Error: False Positive (α)	Correct Decision: (1 - β)

teristics of parametric statistical tests based on Background Test Form 2 is available in Volumes 1 and 2 of the EPA document series *Statistical Methods for Evaluating the Attainment of Cleanup Standards.*[8]

3.2.3 Selecting a Background Test Form

When comparing Background Test Forms 1 and 2, it is important to distinguish between the selection of the null hypothesis, which is a burden-of-proof issue, and the selection of the appropriate level of concern, which involves determination of a quantitative value for a substantial difference based on risk assessment or an action level.

Background Test Form 1 uses a conservative action level of $\Delta = 0$, but relaxes the burden of proof by selecting the null hypothesis that the contaminant concentration in contaminated areas is indistinguishable from background. Background Test Form 2 requires a stricter burden of proof, but relaxes the action level from 0 to S. Section 5.4 includes further discussion of how to choose between Test Forms 1 and 2, and gives additional guidance for setting up the hypotheses.

Regardless of the choice of hypothesis, an incorrect conclusion could be drawn from the data analysis using either form of the test. To account for this inherent uncertainty, one must specify the limits on the Type I and Type II decision errors. This task is addressed in Step 6 of the DQO process and described in the following section.

3.3 Errors Tests and Confidence Levels

A key step in developing a sampling and analysis plan is to establish the level of precision required of the data.[1] Whether the null hypothesis (Section 3.2) will be rejected or not depends on the results of the sampling. Due to the uncertainties in the data, decisions made using the test will be subject to errors. Decisions need to be made about the width of the gray region and degree of decision error that is acceptable. These topics are discussed below and in more detail in Chapter 5. There are two ways to err when analyzing data (see box above):

- *Type I Error*: Based on the data observed, the test may reject the null hypothesis when in fact the null hypothesis is true (a false positive). This is a *Type I error*. The probability of making a Type I error is α (*alpha*); and

- *Type II Error*: On the other hand, the test may fail to reject the null hypothesis when the null hypothesis is in fact false (a false negative). This is a *Type II error*. The probability of making a Type II error is β (*beta*).

The *error tolerance* associated with hypothesis testing is defined by two key parameters—*confidence level* and *power* (see box on next page). These parameters are closely related to the two error probabilities, α and β.

- *Confidence level (1 - α)*: The confidence level for a statistical test is defined as one hundred percent minus alpha. As the confidence level is

Interpretation of the Statistical Measures

Background Test Form 1

Confidence level = 80%: In at least 80 out of 100 cases, site-related chemical concentrations would be correctly identified as being no different (statistically) from background concentrations, while in at most 20 out of 100 cases, site-related concentrations would be incorrectly identified as being greater than background concentrations.

Power = 90%: In 90 out of 100 cases, site-related contaminants would be correctly identified as being greater than background concentrations, while in 10 out of 100 cases, site-related concentrations would be incorrectly identified as being less than or equal to background concentrations.

Background Test Form 2

Confidence level = 90%: In at least 90 out of 100 cases, site-related concentrations would be correctly identified as exceeding background concentrations by more than S, while in at most 10 out of 100 cases, site-related concentrations would be incorrectly identified as not exceeding background concentrations by more than S.

Power = 80%: In at least 80 out of 100 cases, site-related concentrations would be correctly identified as not exceeding background concentrations by more than S, while in at most 20 out of 100 cases, site-related concentrations would be incorrectly identified as exceeding background concentrations by more than S.

lowered (or alternatively, as α is increased), the likelihood of committing a Type I error increases.

- *Power (1 - β)*: The power of a statistical test is defined as one hundred percent minus beta. As the power is lowered (or alternatively, as β is increased), the likelihood of committing a Type II error increases.

Although a range of values can be selected for these two parameters, as the demand for precision increases, the number of samples and the cost will generally also increase.

Because there is an inherent tradeoff between the probability of committing a Type I or Type II error, a simultaneous reduction in both types can only occur by increasing the number of samples. If the probability of committing a false positive is reduced by increasing the level of confidence of the test (in other words, by decreasing α) the probability of committing a false negative is increased because the power of the test is reduced (increasing β).

For the purposes of this guidance, minimum recommended performance measures are:

- For Background Test Form 1, confidence level at least 80% (α = 20%) and power at least 90% (β = 10%)[9]
- For Background Test Form 2, confidence level at least 90% (α = 10%) and power at least 80% (β = 20%)

When using Background Test Form 1, a Type I error (false positive) is less serious than a Type II error (false negative). *This approach favors the protection of human health and the environment.* To ensure that there is a low probability of Type II errors, a Form 1 statistical test must have adequate power at the right edge of the gray region.

When Background Test Form 2 is used, a Type II error is preferable to committing a Type I error. *This approach favors the protection of human health and the environment.* The choice of hypotheses used in Background Test Form 2 is designed to be protective of human health and the environment by requiring that the data contain evidence of *no substantial contamination*. This approach may be contrasted to the "innocent until proven guilty" approach used in Background Test Form 1.

3.4 Test Performance Plots

During the scoping stage for the development of the sampling plan, the interrelationships among the decision parameters can be visualized using a *test performance plot*. The test performance plot is a graph that displays the combined effects of the decision error rates, the gray area for the decision-making process, and the level of a substantial difference between site and background. In short, it displays most of the important parameters developed in the DQO process.

A test performance plot is used in the planning stages of the DQO process to aid in the selection of reasonable values for the decision error rates (α and β), the MDD, and the required number of samples. Selection of these parameters is usually an iterative process. Trial values of the decision error rates, the location of the gray region, and its width (the MDD) are used to generate initial estimates of the required number of samples and the resulting test performance curve. Adjustments to the inputs are made until a design is achieved that offers acceptable test performance at an acceptable cost.

Exhibit 3.1 illustrates an example of a test performance plot for decision making on a statistical test based on the null hypothesis that the *site does not exceed background* (Background Test Form 1). At the origin of the plot, the true difference between the site and background distributions is zero (Δ). Positive values of the difference between the site and background ($\Delta > 0$) are plotted on the horizontal axis to the right of the origin, negative values ($\Delta <$ 0) to the left. The vertical axis shows the value of the test performance measure, defined as the probability of deciding the site exceeds background. This probability ranges from 0 to 1.0 (0 to 100 percent).

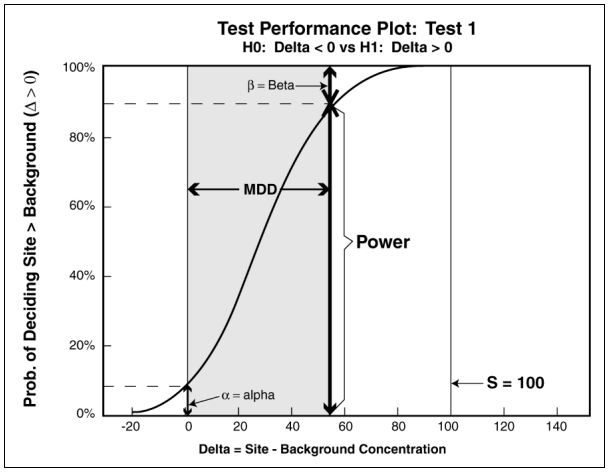

Exhibit 3.1 Test performance plot: site is indistinguishable from background.

At the left edge of the gray region, the test performance curve is no greater than α for contaminated areas with contaminant concentrations less than or equal to background ($\Delta \leq 0$) and greater than α for contaminated areas exceeding background ($\Delta > 0$). The test performance curve increases as the difference between the site and background increases. The number of samples and the standard deviation, σ, determine the rate of increase. The right edge of the gray region is located at the MDD (Δ = MDD). At this value of the difference between the site and background concentrations, the probability of deciding that the site exceeds background is equal to 1 - β. When using Background Test Form 1, the test performance curve equals the power of the test. A statistical software package for plotting the power of a statistical test may be used to generate a test performance plot. EPA has developed two software packages that generate power curves for the two-sample t-test: DEFT[10] and DataQUEST[11].

Exhibit 3.1 shows a hypothetical value of a substantial difference for this chemical of S = 100. The value of S was developed by conducting an evaluation of the risks presented by the site. The value of

S is used in the DQO process as an upper limit for the width of the gray region (MDD). In some cases, an MDD less than S may be selected for the test. This is determined by site-specific conditions, summarized by the standard deviation, σ. If the ratio S/σ exceeds 3, then a sample design with an MDD less than S may offer a test with better power of resolution at little additional cost of sampling, a strategy often described using the term "ALARA"—"As Low As Reasonably Achievable." If the MDD is selected to be smaller than S, then the design is conservative in the sense that sites with differences from background smaller than S can be identified by the test. The test will have a higher power to reject the null hypothesis for sites with mean concentrations that are in the range between the MDD and S higher than background. In statistical terms, the power of rejection will be $(1 - \beta)$ at Δ = MDD, and higher than $(1 - \beta)$ for all $\Delta >$ MDD.

Selecting an MDD less than S is also useful for screening a large number of areas using a low cost sample measurement procedure, with subsequent confirmatory testing using more expensive procedures before making a final decision. Finally, before using previously collected data for decision making, the power of the test should be calculated to determine if the MDD is less that S.

An equivalent plot in Exhibit 3.2 shows the test performance curve for a statistical test using the null hypothesis that *the site does not exceed background by more than a substantial difference* (Background Test Form 2). For this Test Form, the MDD again measures the width of the gray region, but the gray region now extends from a difference of Δ = S-MDD on the left to a difference Δ = S on the right.

When using Background Test Form 2, the MDD may be selected to be as large as S or smaller. The implications of making the MDD smaller than S for this Test Form differ from those that occur when using Background Test Form 1. As the MDD decreases below S, the test will identify more sites as not having mean concentrations that exceed background by more than S. The sites with mean concentration in the range between Δ = 0 and Δ = S - MDD (those with mean concentrations only slightly higher than background) will have a higher probability of being classified correctly. With this Test Form, a tradeoff exists between taking more samples and making more errors. Since the errors tend to occur in sites that are marginally acceptable, site owners/operators have an incentive to increase the number of samples and the power of the test.

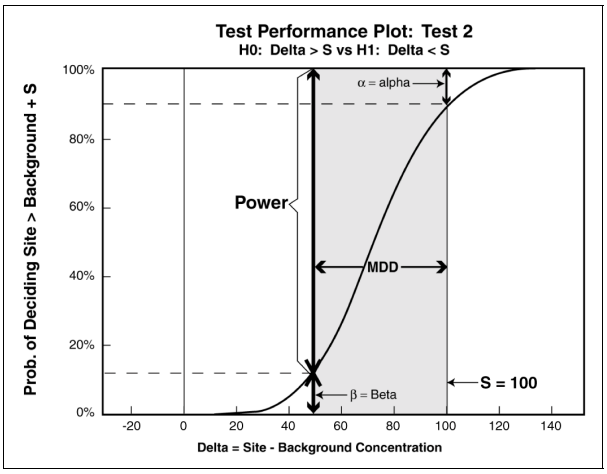

Exhibit 3.2 Test performance plot: site does not exceed background by more than S.

This second form of background test requires switching the location of α and β. The Type I error (α) for Background Test Form 2 is measured by the difference between 100% and the test performance curve at the right of the gray region, while the Type II error (β) is measured by the value of the test performance curve at a difference equal to Δ = MDD, located at the left of the gray region. When using Background Test Form 2, the test performance curve equals 100% minus the power of the test.

Comparison of Exhibits 3.1 and 3.2 demonstrates that the choice $\alpha = \beta$ and MDD = S will result in almost identical test performance plots for Background Test Form 1 and Background Test Form 2. If $\alpha = \beta$ and MDD is less than S, then Background Test Form 1 will indicate that more contaminated areas require remediation than Background Test Form 2. In general, α will differ from β, and the value selected for the MDD may be smaller than S.

When using Background Test Form 1, a Type I error could lead to unnecessary remediation while a Type II error could lead to unacceptable health risks. If Background Test Form 2 is used, a Type II error could lead to unnecessary remediation while a Type I error could lead to unacceptable health risks. Therefore, one should attempt to reduce the chance of making either of these errors.

The selection of tolerable decision error values for hypothesis testing is a decision that must be made on a site-specific basis. The consequences of making a wrong decision (such as failing to reject the null hypothesis when it is false) should be considered when specifying acceptable values for the confidence and power factors (α = 20% and β = 10% are maximum values for Background Test Form 1).

3.5 Sample Size

In most DQO applications, after electing to use a test with confidence level 100(1 - α) percent, the required number of samples is determined by simultaneously selecting:

- the MDD for the test; and
- the power (1 - β) of the test at the MDD.

Therefore, limits on the probability of committing Type I and Type II errors can be used as constraints on the number and location of samples. To determine realistic limits for the decision errors, the number of samples (and the corresponding cost of sampling) could be estimated for a range of probability values, which would indicate the likelihood of making either type of error.

Several reference documents give formulas or tables for selecting the number of samples, given the specific confidence and power limits.[12] Chapter 5 offers guidance for selecting appropriate statistical techniques for comparing on-site and background contaminant concentrations in soil.

Examples of constraints that may be adjusted to influence the required sample size include:

- Increasing the decision error rates, α and β, while considering the increased costs and risks associated with the increased probability of making an incorrect decision;

- Increasing the width of the gray region (MDD), but do not exceed a substantial difference (MDD $\leq$ S); and

- Changing the boundaries. It may be possible to reduce measurement costs by segregating the site into subunits that require different decision parameters due to different risks.

The site team should consult with a statistician to select the appropriate sampling design. Several sampling design options are available. A consistent grid to cover the entire site and areas considered as background should provide a reasonable characterization of the contamination and background. An alternative design option might involve collection from areas where site contamination appears homogenous. Site history and past activities could be used to vary grid size and intensify sampling efforts in the potentially contaminated areas versus areas with little or no past activity. Additional options are described in other guidance, including Chapter 4 of RAGS.

3.6 An Example of the DQO Process

This section introduces an example application of the DQO process for comparing lead concentrations in a contaminated area to background. This is a hypothetical example. The conceptual site model and remedial goals for individual sites will determine what sampling and analysis is done at any site. The example will illustrate some outputs of the DQO process and will be extended to the preliminary data analysis stage in Chapter 4 and to the hypothesis testing stage in Chapter 5. The Superfund Program has a Technical Review Workgroup for Lead (TRW) that can provide technical assistance.[13]

Step 1. State the Problem

An abandoned storage yard has been identified as the possible source of elevated lead levels found in neighborhood wells. Although other sources of background lead are present in the vicinity of the storage yard due to nearby highways and industrial facilities, concerns about the site have focused on the storage area as a likely source of contamination. Also, the available data are not sufficient to determine that the site concentrations are different from background chemical concentrations. An investigator has been assigned to conduct field measurements. Tasks include:

a. *Identify the resources available to resolve the problem, including scoping team*

The members of the planning team will include the plant manager, a plant engineer, a chemist with field sampling experience, a quality assurance officer, a statistician, and the investigator assigned by the EPA.

b. *Develop or refine the comprehensive conceptual site model*

Historical site assessment was used to develop a comprehensive conceptual site model. Due to nearby highways and industrial sources in the vicinity of the yard, background lead concentrations in soil are expected to be above the national average. Because of run-off from paved areas, background concentration near paved areas are likely to be higher than background concentrations in soils that are distant from paved areas. The selection of appropriate background areas for the comparison was restricted to areas at least 1,000 meters from heavily used highways and 30 meters from paved surfaces.

Step 2. Identify the Decision

Do soils in the storage area have higher lead contamination than found in soils in the surrounding area, and if so, are they attributable to the storage area? Tasks include:

a. *Identify the chemicals to analyze*

The purpose of the study is to compare total lead concentrations at the storage yard and in surrounding background areas.

b. *Determine if the chemical is likely to be a background constituent.*

Because of the nearby highways and other industrial sources in the vicinity of the yard, background lead concentrations are expected to be elevated. Background concentrations near paved areas are likely to be higher than background concentrations that are far distant from paved areas.

Step 3. Identify Inputs into the Decision

a. *Which chemicals will be analyzed?*

The study team decides to focus on total lead concentration.

b. *Which soil types and depths need to be sampled?*

Because there is neither surface evidence, nor historical record, of excavation in the storage area, the study team decides to measure total lead concentration in the first 12 inches of surface soils. Soils in background locations will be sampled in the same way. The TRW has recommended soil sieving at 250 µm to assess exposures to lead on the fine fraction of soil and dust.[14] For background sampling of lead, this fractionation may be appropriate as it relates to human health risks. An average soil sample depth is reasonable, but dust samples naturally will be collected in shallow or surficial layers.

c. *Which comparison tests are likely to be used?*

The study team expects that lead concentrations may not be normally or lognornally distributed. Although the data may permit use of more powerful tests if these distributions do apply, the study team decides to use a nonparametric statistical test for differences in the soil lead concentration distribution in the

storage yard and in the surrounding areas.

d. *What coefficient of variation is expected for the data?*

Based on previous sampling in other areas, a coefficient of variation ranging from 50% to 200% is expected. The team expects that stratification of the site into paved and unpaved areas will reduce the variability within each stratum. The team decides that a coefficient of variation of 100% is expected within a stratum. The study team agrees to review this decision, depending on the overall cost estimates produced by the decision objectives.

e. *What preliminary remediation goals (PRGs) may need to be met?*

A PRG of 400 mg/kg is available for residential sites.[15]

f. *Specify the desired power and confidence levels*

The study team decides initially on a Type I decision error limit of α = 10% and a Type II decision error limit of β = 10% (power = 90%). The team agrees to review this decision, depending on the overall cost estimates produced by these objectives.

Step 4. Define Boundaries of the Study

a. *Define the geographic areas for field investigation*

The study team decides that the entire storage yard area, approximately 5 acres, will be included in the study. Four different background areas of approximately 10,000 m^2 were selected at distances of between 1,000 m and 10,000 m from the storage yard boundaries.

b. *Define the characteristics of the soil data or population of interest*

Soil samples should be collected in dry, unpaved areas. Prepared samples should be free of roots, leaves, and rocks or other consolidated materials. When preparing the samples, these materials should be removed using a 3 cm diameter sieve. Oversized materials will be retained for additional weighing and analysis, if necessary.

c. *Divide the soil data population of interest into strata having relatively homogeneous characteristics*

Stratification of the site data into paved and unpaved is planned for this sampling.

d. *Determine the time frame to which the decision applies*

Sampling will be conducted during a four-week period in the fall. Lead concentrations in soil are relatively static, and decisions based on the sampling results will remain applicable for many years, barring additional contamination.

e. *Identify practical constraints that may hinder sample collection*

The plant manager agreed to permit EPA sampling on the storage yard. Permission must be obtained from the owners of the selected background sampling areas for permission to enter and to collect background samples on their property.

Step 5. Develop a Decision Rule

If the selected statistical test indicates that the mean concentration in contaminated areas exceeds the mean background concentration, then the chemical will be treated as site-related. Otherwise, if the statistical test indicates that the mean concentration in contaminated areas does not exceed the background mean, the chemical will be treated as coming from the same population as background.

a. *Choose the null hypothesis*

The study team chooses a null hypothesis that the lead concentrations in the storage yard exceed background concentrations.

- H_0: Lead concentrations in the storage yard samples exceed background concentrations by more than S = 50 mg/kg (see paragraphs c and d, below, for how 50 mg/kg was chosen).

b. *Specify the alternative hypothesis*

The alternative hypothesis is the opposite of the null hypothesis.

- H_A: Lead concentrations in the storage yard samples do not exceed the background concentrations by more than S = 50 mg/kg.

c. *Determine the level of a substantial difference above background*

The study team has decided to use a value of 100 mg/kg as the value for a substantial difference in lead concentrations between the storage yard and background areas. This decision was based on the fact that EPA remedial goals for residential soils for lead contamination often range from 400 mg/kg to about 1,000 mg/kg.[16] The selected value of S represents less than 10% of the higher end of this range of remedial goals.

d. *Specify the gray region for the hypothesis test*

When using Background Test Form 2, the gray region of width MDD starts at a difference of Δ = S = 100 mg/kg and extends on the left down to Δ = (S - MDD). As a trial value, the study team choose to use an MDD that is one-half of S, 50 mg/kg (refer to Exhibit 3.3).

The site manager conducted a trade-off analysis between the cost of extra sampling and the expected cost of remediating the site unnecessarily, and decided to make the width of the gray region one-half of S (refer to Exhibit 3.3).

Test Form 2 has at least 100(1-α)% confidence of correctly detecting a site that exceeds background by more than S, regardless of the sample size. Greater sample size increases the power of the test and reduces β, which reduces the chance that a site is remediated unnecessarily. When using Test Form 2, extra samples represent the cost of increasing the chance that the site is accepted when the true Δ is less than S. The study team agrees to review this decision, depending on the overall cost estimates produced by the decision objectives.

σ	MDD/σ (mg/kg)	n
25	2	5
50	1	16
75	0.67	36
100	0.50	63
125	0.40	98
150	0.33	140
175	0.29	190
200	0.25	248

Exhibit 3.3 Required sample size for selected values of σ

Step 6. Specify the Limits on Decision Errors

a. *Determine the possible range of the parameter of interest*

The possible range of lead concentrations in industrial soil is very wide, ranging from 0 to many grams per kilogram.

b. *Specify both types of decision errors (Type I and Type II)*

The team decides that the acceptable limits on decision errors are α = 10% for Type I errors at a difference of Δ = S = 100 mg/kg, and β = 10% for Type II errors at a difference of Δ = S/2 = 50 mg/kg.

The investigator was comfortable with the choice of a 90% confidence level (α=0.10) of the test, because this reduces the chance of a false negative—deciding that the yard does not exceed background by more than S. In Exhibit 3.2, the test performance curve achieves a probability of 90% of detecting a site at Δ = S. This reduces the probability of a false negative to 10% at a difference of Δ = S = 100

mg/kg, and to less than 10% at a higher value of Δ.

The choice of $\beta = 10\%$ and the selected value for the MDD equal to one-half the width of the gray region means that the power of 90% will be required at $\Delta = S/2$. The plant manager recognizes that a lower value of β (higher power) would result in a lower probability of a Type II error and improve his chances of passing the test, but he has decided during the trade-off analysis in Step 5d that the extra sampling costs required to achieve a higher power are not necessary.

c. *Identify the potential consequences of each type or error, specifying a range of possible parameter values (gray area) where consequences of decision errors are relatively minor*

The team decides that the decision errors are $\alpha = 10\%$ at $\Delta = S$, and $\beta = 10\%$ at $\Delta = S/2$. The gray region extends from a difference $\Delta = 50$ mg/kg to a difference of $\Delta = 100$ mg/kg (refer to Exhibit 3.2 and decisions made in Steps 5d and 6b).

d. *Check the limits on decision errors to ensure that they accurately reflect the study team's concern about the relative consequences for each type of decision error*

The investigator is satisfied with the choice of the 90% confidence level for the statistical test, because this will reduce to 10% the chance of falsely deciding that the yard does not exceed background by more than 100 mg/kg when it truly does. The use of a level-α test will provide 90% confidence for all sample sizes, but may have poor power if the sample size is too low.

The sample size is fixed by the choice of MDD and β. Choosing $\beta = 10\%$ at a difference of $\Delta = 50$ mg/kg means that a power of at least 90% will be obtained if the true lead concentration on the yard is at or below that value. The plant manager recognizes that a lower value of β (higher power) would result in a lower probability that the test will decide the yard exceeds background lead concentrations if the yard is only 50 mg/kg higher than background. However, the manager has decided that this extra power would require more sampling and unwanted additional sampling costs.

The DQO parameters α, β, S, and MDD provide almost all that is needed to calculate the number of samples (N) required from each population. Thus, N samples will be collected in contaminated areas, and a total of N samples will be collected in the background locations.

The only remaining parameter required is an estimate of the standard deviation of the soil lead concentrations in contaminated areas (σ). Since the variability is usually higher in the contaminated areas than in background locations, the standard deviation in contaminated areas is used to estimate the required sample size. The estimate for σ usually is obtained from historical data, if available. Alternatively, estimates of variability reported elsewhere at similar sites with similar contamination problems may be used. If an estimate of the mean concentration in contaminated areas is available, the coefficient of variation observed at other sites may be multiplied by the mean to estimate the standard deviation.[17]

The sample size may be calculated using the approximate formulas presented in Chapter 3 of EPA QA/G9.[2] More specific sample-size calculation procedures are given in a multi-agency manual.[18] Sample sizes obtained using the approximate formulas in EPA QA/G9 are shown in Exhibit 3.3 for a variety of σ values. Note that the required sample size increases dramatically when the MDD is smaller than $\sigma/2$.

A general rule of thumb to obtain a reasonable sample size is to set the MDD approximately equal to σ.

Step 7. Optimize the Sampling Design

What is the most resource effective sampling and analysis design for generating data that are expected to satisfy the DQOs?

a. *Review the DQO outputs and existing environmental data*

The statistician, chemist, and plant engineer on the study team have reviewed the outputs developed at each stage of the DQO process.

b. *Develop general sampling and analysis design alternatives*

The study team decides to use a randomly-oriented, rectangular grid sampling strategy for the storage yard and selected background area. Two random numbers (x and y) randomly will determine the starting point selected for the grid. The grid orientation will be determined by a third random number. The size of the grid will be calculated based on the number of samples required for each area.

c. *Verify that DQOs are satisfied for each design alternative*

Only one sample design is used in this study.

d. *Select the most resource-effective design that satisfies all of the DQOs*

The stratified design will reduce variability within the strata, resulting in lower sampling costs.

e. *Document the operational details and theoretical assumptions of the selected design in the sampling and analysis plan*

The study team has documented the discussions leading to each DQO parameter.

CHAPTER NOTES

1. U.S. Environmental Protection Agency (EPA). 1994. *Guidance for the Data Quality Objectives Process, EPA QA/G-4*, EPA 600-R-96-065. Washington DC.

2. U.S. Environmental Protection Agency (EPA). 2000. *Guidance for Data Quality Assessment: Practical Methods for Data Analysis, EPA QA/G-9, QA00 Version*. Quality Assurance Management Staff, Washington, DC, EPA 600-R-96-084. Available at http://www.epa.gov/quality/qa_docs.html.

3. U.S. Environmental Protection Agency (EPA). 1989. *Risk Assessment Guidance for Superfund Vol. I, Human Health Evaluation Manual (Part A).* Office of Emergency and Remedial Response, Washington, DC. EPA 540-1-89-002. Hereafter referred to as "RAGS."

4. Mathematically, Background Test Form 1 is written:
   $$H_0: \Delta \leq 0 \text{ } vs \text{ } H_A: \Delta > 0$$
   with $\Delta = \theta_S - \theta_B$, where θ_S is the selected decision parameter (mean, median, etc.) for the site distribution, and θ_B is the same parameter for the background distribution.

5. Equations for computing retrospective power are provided in the detailed step-by-step instructions for each hypothesis test procedure in Chapter 3 of *Guidance for Data Quality Assessment: Practical Methods for Data Analysis, EPA QA/G-9, QA00 Version. Op.Cit*.

6. U.S. Environmental Protection Agency (EPA). 1989. *Statistical Methods for Evaluating the Attainment of Cleanup Standards*, EPA 230/02-89-042, Washington DC.

7. Mathematically, Background Test Form 2 uses the substantial difference S as a non-zero action level:
   $$H_0: \Delta > S \text{ } vs \text{ } H_A: \Delta \leq S$$
   with $\Delta = \theta_S - \theta_B$, where θ_S is the selected decision parameter (mean, median, etc.) for the site distribution, and θ_B is the same parameter for the background distribution.

8. U.S. EPA. 1989. *Statistical Methods for Evaluating the Attainment of Cleanup Standards*. *Op. cit.*

9. U.S. Environmental Protection Agency (EPA). 1990. *Guidance for Data Usability in Risk Assessment: Interim Final, October 1990*. EPA 540-G-90-008, PB91-921208, Washington, DC.

10. U.S. Environmental Protection Agency (EPA). 1994. *The Data Quality Objectives Decision Error Feasibility Trials (DEFT) Software (EPA QA/G-4D)*, EPA/600/R-96/056, Office of Research and Development, Washington, DC.

11. U.S. Environmental Protection Agency (EPA). 1996. *The Data Quality Evaluation Statistical Toolbox (DataQUEST) Software (EPA QA/G-9D)*, Office of Research and Development, Washington, DC.

12. Common references for sample selection include:

    - Cochran, W. 1977. *Sampling Techniques*. New York: John Wiley.

- Gilbert, Richard O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. New York: Van Nostrand Reinhold.

- U.S. EPA. 1989. *Statistical Methods for Evaluating the Attainment of Cleanup Standards*. *Op. cit.*

- U.S. EPA, 1990. *Guidance for Data Usability in Risk Assessment*. *Op. cit.*

13. The Technical Review Workgroup for Lead provides technical assistance for people working on lead-contaminated sites. For assistance or more information, the reader should refer to their website (http://epa.gov/superfund/programs/lead) or call the Lead Hotline (800-680-5323).

14. U.S. EPA. 2000. *TRW Recommendations for Sampling and Analysis of Soil at Lead (Pb) Sites.* Office of Emergency and Remedial Response, Washington, DC. EPA 540-F-00-010, OSWER 9285.7-38.

15. U.S. EPA. 1994. *Revised Interim Soil Lead Guidance for CERCLA Sites and RCRA Corrective Action Facilities.* OSWER Directive 9355.4-12.

16. U.S. Environmental Protection Agency (EPA). 1995. *BESCORP Soil Washing System for Lead Battery Site Treatment.* EPA 540-AR-93-503. Office of Research and Development, Washington, DC.

17. If no acceptable source for an estimate of σ is available, it may be necessary to conduct a small-scale pilot survey on site using 10 or more random samples to estimate σ. Due to the small sample size of the pilot, it is advisable to use an 80 or 90 percent upper confidence limit for the estimate of σ rather than an unbiased estimate to avoid underestimating the true variability. A very crude approximation for σ may be made by dividing the anticipated range (maximum - minimum) by 6.

18. U.S. Environmental Protection Agency (EPA), U.S. Nuclear Regulatory Commission, et al. 2000. *Multi-Agency Radiation Survey and Site Investigation Manual (MARSSIM)*. Revision 1. EPA 402-R-97-016. Available at http://www.epa.gov/radiation/marssim/ or from http://bookstore.gpo.gov/index.html (GPO Stock Number for Revision 1 is 052-020-00814-1).

CHAPTER 4

PRELIMINARY DATA ANALYSIS

This chapter provides guidance for preliminary data analysis using graphs and distributions of the data. Depending upon the quality of existing data on site contamination, quantitative analysis used to establish background concentration may involve a combination of comparative statistical analysis and graphical methods. The preliminary data analysis is an integral part of choosing the appropriate methods for making meaningful background comparisons to site contamination.

Preliminary data analysis should include a detailed "hands-on" inspection of the site and background data before proceeding to the statistical tests. The preliminary inspection may include development of a posting plot,[1] which is a map showing the location of each sample. The posting plot may reveal likely sources of contamination, important areas that have not been sampled, spatial correlations or trends in the data and the location of suspected outliers. Note that one possible outcome of the preliminary data inspection is that the contaminant concentrations on site greatly exceed background levels, making formal statistical comparisons unnecessary.

This chapter presents information useful for both parametric and nonparametric data analysis. Most parametric statistical methods are based on the assumption of a known mathematical form for the probability distributions that represent the site and background populations. For many parametric methods, the data user must first determine whether the data are normally distributed, using any of several tests for normality.

Nonparametric methods do not require that the data distribution be characterized by a known family of distributions. Several graphical methods are available for nonparametric comparisons.

Parametric and Nonparametric Methods

Parametric: A statistical method that relies on a known probability distribution for the population from which the data are selected. Parametric statistical tests are used to evaluate statements (hypotheses) concerning the parameters of the distribution.

Nonparametric: A distribution-free statistical method that does not depend on knowledge of the population distribution.

Data analysis can encompass either the whole data set from the site, focus on outliers in the background data set, or emphasize site contaminant concentrations. These topics are discussed in the following sections.

4.1 Tests for Normality

Tests should be conducted on each data set to show whether it meets the assumption of normality. If the raw data are not normally or lognormally distributed, other types of transformations should be conducted to approximate normality prior to using the data sets in parametric statistical comparisons, such as t-tests or the analysis of variance procedure (ANOVA). The assumption of normality is very important as it is the mathematical basis for the majority of statistical tests. Examples of how to perform each of these tests can be found in Chapter

4 of EPA's *Guidance for Data Quality Assessment*. The *Shapiro-Wilk* test is a powerful general purpose test for normality or lognormality when the sample size is less than or equal to 50, and is highly recommended. The Shapiro-Wilk test is an effective method for testing whether a data set has been drawn from an underlying normal distribution. It can also evaluate lognormality if the test is conducted on logarithms of the data. If the normal probability plot is approximately linear—the data follow a normal curve—the test statistic will be relatively high. If the normal probability plot contains significant curves, the test statistic will be relatively low.

Another test related to the Shapiro-Wilk test is the *Filliben statistic*, also called the "probability plot correlation coefficient." If the normal probability plot is approximately linear, the correlation coefficient is relatively high. If the normal probability plot contains significant curves—the data do not follow a normal curve—the correlation coefficient will be relatively low. The Filliben test is recommended for sample sizes less than or equal to 100.

D'Agostino's test for normality or lognormality is used when sample sizes are greater than 50. This test is based on an estimate of the standard deviation obtained using the ranks of the data. This estimate is compared to the usual mean square estimate of the standard deviation, which is appropriate for the normal distribution.

The *studentized range test* for normality is based on the fact that almost 100 percent of the area of a normal curve lies within ± 5 standard deviations from the mean. The studentized range test compares the range of the sample to the sample standard deviation. For example, if the minimum of 50 data points is 40.2, the maximum is 62.7 and the standard deviation is 4.2, then the studentized range is (62.7 - 40.2)/4.2 = 5.4. Tables of critical sizes up to 1,000 are available for determining whether the absolute value of the studentized range is significantly large. The studentized range test does not perform well if the data are asymmetric and if the tails of the data are heavier than the normal distribution. In most cases, this test performs as well as the Shapiro-Wilk test and is easier to apply.

4.2 Graphing the Data

Graphical methods provide visual examination of the site and background distributions, and comparisons of the two. Graphical methods supplement the statistical tests described in Chapter 5. Graphical methods also may be used to verify that the assumptions of statistical tests are satisfied, to identify outliers, and to estimate parameters of probability distributions fit to the data.

4.2.1 Quantile Plot

A *quantile plot* displays the entire distribution of the data, ranging from the lowest value to the highest value. The vertical axis for the quantile plot is the measured concentration, and the horizontal axis is the percentile of the distribution. Each ranked data value is plotted against the percentage of the data with that value or less.

To construct a quantile plot, the data set is ranked from smallest to largest. The percentage value for each data point j is computed as

$$Percent_j = 100 (rank_j - 0.5) / n$$

where n is the number of values in the data set. If one or more data values are non-detects, all non-detects are ranked first, below the first numerical value. The plot starts with the first numerical value.

The slope of the curve in the quantile plot is an indication of the amount of data in a given range of values. A small amount of data in a given range will result in a large slope for the quantile plot. A large amount of data in a range will result in a more horizontal slope. A sharp rise near the bottom or the top of the curve may indicate the presence of outliers.

A graph may contain more than one quantile plot. In a *double-quantile plot* the site and background data are each plotted in a single graph, providing a direct visual comparison of the two distributions. A curve

that is higher in the vertical direction indicates a higher distribution of data values.

An example of the double-quantile plot is shown in Exhibit 4.1. The lower curve shows the distribution of the background data, and the middle curve (indicated by symbols only) shows the quantile plot for the site data. In this example, the entire site distribution is higher than the background distribution indicating that some degree of contamination is likely. The close proximity of the site and background quantile plots near the 70th percentile indicates that the two distributions differ mainly in the upper 30 percent of the distributions.

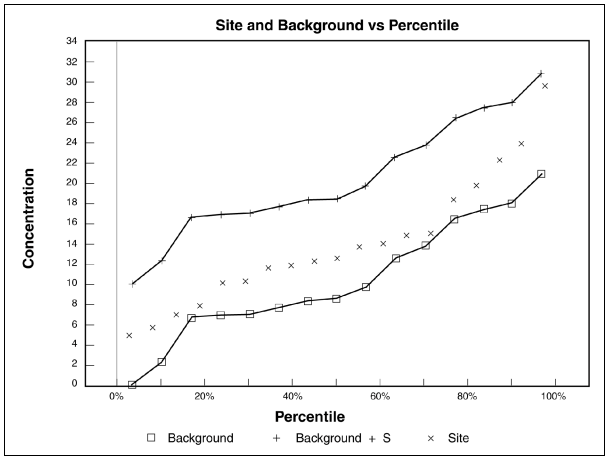

Exhibit 4.1 Example of a double quantile plot.

The upper curve in the exhibit shows the background distribution augmented by S = 10, a hypothetical value for a substantial difference over background. In this example, the entire site distribution lies below the S-augmented background distribution, indicating that the site does not exceed background by more than a substantial difference.

Issues affecting the determination of site-specific values for a substantial difference are discussed in more detail in the Appendix at the end of this guidance.

The formal statistical test procedures presented in Chapter 5 may be used to make decisions that confirm or deny these graphical indications with predetermined error rates. In this and the following exhibits, contaminant concentrations are plotted using a linear scale. If the data are highly variable, it may be necessary to transform the graph by using a logarithmic scale for the concentration axis. Use of the logarithmic transformation does not affect the ranks of the data.

4.2.2 Quantile-Quantile Plots

A *quantile-quantile plot* is useful for comparing two distributions in a single graph. The vertical axis of this plot represents the first distribution of values, and the horizontal axis represents the second distribution. The scales for the concentration axes may be either both linear or both logarithmic. If the two distributions are identical, the quantile-quantile plot will form a straight line at 45 degrees when equal scales are used for the two axes. The slope of this line has a value of one, regardless of the selected scales. Deviations from this line show the differences between the two distributions.

There are two common applications of the quantile-quantile plot. One type is used for parametric applications, and the other for nonparametric comparisons.

- *Parametric Quantile-Quantile Plot*. In parametric applications of the quantile-quantile plot, the horizontal axis represents the quantiles from a known distribution, such as the normal distribution. This application is referred to as a *normal probability plot.* If the data follow a normal distribution, the plot will appear as a straight line. Probability plots are useful for determining if the site data or the background data follow a normal or lognormal distribution. More information on the use of the quantile-quantile plot to compare with known parametric distributions is provided in EPA's *Guidance for Data Quality Assessment*, Section 2.3.

- *Empirical Quantile-Quantile Plot*. In nonparametric applications, the empirical quantile-quantile plot is used to compare two data sets. In our case, the two data sets are the site distribution and the background distribution. If

there are an equal number of data values in the two data sets, it is very easy to construct an empirical quantile-quantile plot. The graph is constructed by plotting each ranked site value against the corresponding background value with the same rank. When the size of the site data set differs from the size of the background data set, interpolation is used to construct the empirical quantile-quantile plot. The interpolation method is discussed below.

The empirical quantile-quantile plot is useful because it provides a direct visual comparison of the two data sets. An example of the quantile-quantile plot is shown in Exhibit 4.2. If the site and background distributions are identical, the plotted values would lie on a straight line through the origin with slope equal to 1, shown in the exhibit as the line labeled "Site=Background." Any deviation from this line shows differences between the two distributions. If the site differs from the background data distributions only by an additive difference along the entire distribution, the plotted site values will lie on a straight line with slope 1 that does not pass through the origin. If the site distribution is t units above the background distribution, the straight line will have slope 1 and a y intercept at + t.

A hypothetical level of substantial contamination, S, is shown in the upper plot in Exhibit 4.2 labeled "Background + S." Note that the median interpolated site value is plotted against the median of the background values at the center of the plot. When this point lies above the equal-distribution line with slope 1, the median interpolated site value is larger than the median background value.

In the more likely case, the site data set will have a different number of data values than the background data set. To construct an empirical quantile-quantile plot in this case, interpolation is used to calculate a value from the larger data set that corresponds with each ranked value in the smaller data set. Detailed procedures for creating a quantile-quantile plot with unequal sample sizes are provided in Section 2.3.7.4 of EPA QA/G9.[1]

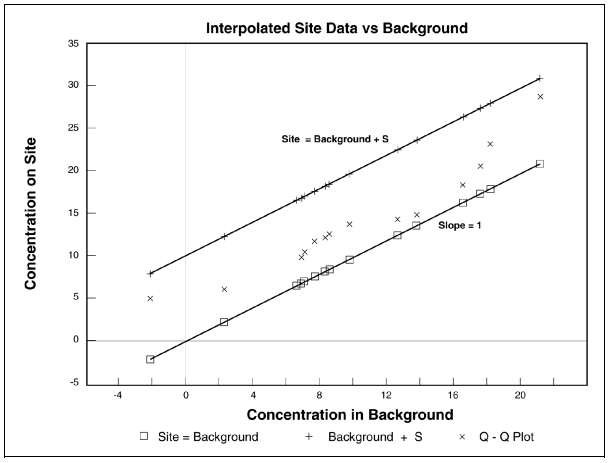

Exhibit 4.2 Example of a quantile-quantile plot.

4.2.3 Quantile Difference Plot

The *quantile difference plot* is a variant of the empirical quantile-quantile plot. When site data are compared to background data, the quantity of greatest interest is the amount by which the site distribution exceeds the background distribution. This difference can be viewed in the quantile-quantile plot as the difference between two sloped lines, the quantile-quantile plot and the line with slope 1 where site equals background. More resolution for examining the differences between the site and background distributions is obtained by subtracting each background value from its corresponding interpolated site value, then plotting the differences versus their corresponding background values.

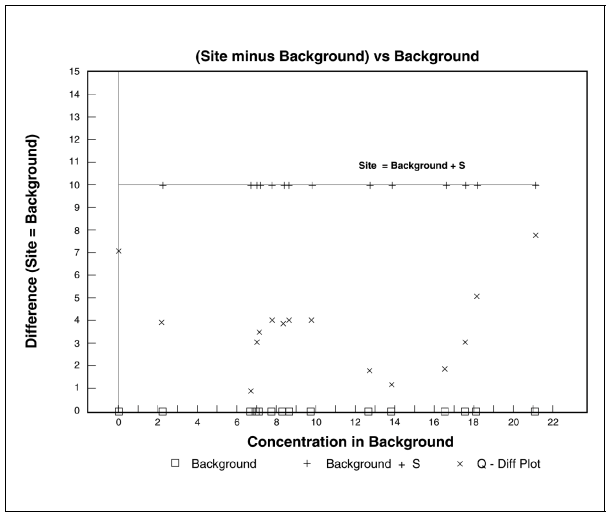

Exhibit 4.3 Example of a quantile difference plot.

An example of the quantile difference plot is shown in Exhibit 4.3. In the quantile difference plot, background is represented by the horizontal axis. The distribution of background values is shown by the symbols plotted on this axis. A hypothetical level of substantial contamination of S = 10 appears in this plot as a horizontal line, not to be exceeded. In this example, the entire quantile difference plot lies between the background and the substantial difference level, indicating that the site exceeds background by a small amount, but does not exceed background by more than a substantial difference.

The quantile difference plot permits a quick visual evaluation of the amount by which the site exceeds background. In this example, the largest differences occur in the upper half of the distribution. It is clear that the interpolated site values do not exceed background by more than the hypothetical S = 10 concentration units. This conclusion is not as obvious using the sloped nonparametric quantile-quantile plot.

Similar warnings exist for use of the quantile difference plot as for the empirical quantile-quantile plot when there are more than twice as many site values as background values. The empirical quantile-quantile plot and the quantile difference plot work best when the site and background data sets are of approximately the same size, and they depend upon the choice of S.

4.3 Outliers

Outliers are measurements that are unusually larger or smaller than the remaining the data. They are not representative of the sample population from which they were drawn, and they distort statistics if used in any calculations. Statistical tests based on parametric methods generally are more sensitive to the existence of outliers in either the site or background data sets than are those based on nonparametric methods.

Outliers can lead to both Type I and Type II errors. They can lead to inconclusive results if the results are highly sensitive to the outliers.

There are many plausible reasons for the presence of outliers in a data set:

- Data entry errors. Data that are extremely high or low should be verified for data entry errors.

- Missing values and non-detects. It is important that missing value and non-detect codes are not read as real data. For example the number 999 might be a code for missing data but the computer program used to analyze the data, if not properly designated, could misread this as an extreme value of 999. This is easily remedied.

- Sampling error. In this case the sample results for the sample that is not from the population of interest should be deleted. However, using data from a population other than the one of concern is not easily recognized. Therefore, this type of error can result in the presence of outliers in the data set.

- Non-normal population. An outlier might also exist when a sample is from the population of interest, but its distribution has more extreme values than the normal distribution. In this situation, the sample can be retained if a robust statistical approach is selected so that the outliers do not have undue impact.

Outliers may misrepresent the sample population from which they were taken, and any conclusion drawn that is based on these results may be suspect. If there is a large number of outliers in the data set, it may be necessary to reassess the area. Outliers in the site data set have different implications from outliers in the background data set. For example, an on-site outlier can indicate a “hot spot”—the contaminant passes the t-test, but one sample exceeds the upper tolerance limit (UTL), which indicates that the one spot needs attention. An outlier in the background data set, however, might indicate that one of the background samples was collected in a location that is not truly background.

In such a case, an outlier test should be used (along with a qualitative study of where the sample in question was collected) to see if that data point should be discarded from the background set.

Statistical outlier tests give probabilistic evidence that an extreme value does not "fit" with the distribution of the remainder of the data and is therefore a statistical outlier. There are five steps involved in treating extreme values or outliers:

1. Identify extreme values that may be potential outliers;

2. Apply statistical tests;

3. Scientifically review statistical outliers and decide on their disposition;

4. Conduct data analyses with and without statistical outliers; and

5. Document the entire process.

More guidance on handling outliers is given in Chapter 5.

4.4 Censored Data (Non-Detects)

Contamination on the site or in background areas may be present at concentrations close to the detection limits. A sample is said to be "censored" when certain values are unknown, although their existence is known. *Type I censoring* occurs when the sample is censored by reference to a fixed value. Non-detect measurements are examples of Type I left censoring. The specific value is unknown, but the existence of a concentration value in the interval from 0 to the reporting limit is known. Concentration values may be censored at their detection limits or at some arbitrary level based on detection limits.

A detection limit is the smallest concentration of a substance that can be distinguished from zero. Consequently, non-detects may not represent the absence of a chemical but its presence at a concentration below its reliable minimum detection level. Many parametric statistical methods require numerical values for all data points. One approach is to impute a surrogate value for non-detects, commonly assumed to be half the reporting limit. As an alternative to half the reporting limit, a random value between the reporting limit and zero may be chosen to represent each non-detect for the purposes of testing assumptions concerning distributions. Both approaches may seriously affect the estimated distribution parameters.

If less than 50 percent of the site and background samples are non-detects, then distributions of both the background and the site sample may be determined by using surrogate values. Probability plots and goodness-of-fit tests may be performed for each data set, first including the non-detects as part of the sample using random values for non-detects, and second, excluding the non-detects from the sample. If the two sets of estimated parameters differ only slightly, then the non-detect problem is of lesser importance. However, if the two sets of estimates differ significantly, then the surrogate value approach should be re-evaluated.

If more than 50 percent of the measurements in either the background sample set or the site sample set are non-detects, it may not be possible to compare the means of the two distributions. An alternative approach is to compare the upper percentiles of the two distributions by comparing the proportion of the two populations that is above a fixed level, as recommended in EPA QA/G-9.[2] Comparisons may be made for the upper percentiles of each distribution despite the large number of non-detects.

Nonparametric methods may be used to avoid the necessity of imputing surrogate values for non-detect measurement. Nonparametric methods are often based only on the ranks of the data, and the non-detect values can be assigned unambiguous ranks without the need for assigning surrogate values.

CHAPTER NOTE

1. U.S. Environmental Protection Agency (EPA). 2000. *Guidance for Data Quality Assessment: Practical Methods for Data Analysis, EPA QA/G-9, QA00 Version*. EPA 600-R-96-084. Quality Assurance Management Staff, Washington, DC. Available at http://www.epa.gov/quality/qa_docs.html. See Section 2.3.9.1 for guidance on preparing a posting plot.

2. *Ibid*, Section 3.3.2.1.

CHAPTER 5

COMPARING SITE DATA TO BACKGROUND DATA

This chapter provides guidance on selecting quantitative statistical approaches for comparing site data to background data. Statistical methods allow for specifying (controlling) the probabilities of making decision errors and for extrapolating from a set of measurements to the entire site in a scientifically valid fashion.[1]

Several methods are available for comparing background to site data. These can be divided into several major categories: data ranking and plotting, descriptive summaries, simple comparisons, parametric tests, and nonparametric tests. For many of these methods, data users first must determine whether the data are normally distributed, using any of several tests for normality. Data can also be assessed in terms of the whole data set from the site, or with a focus on outliers in the background data set or in the contaminant concentrations at the site (see Chapter 4).

The issue of randomness is an important element of most statistical procedures when sample results are to be extrapolated to the entire site or background sampling area, rather than only representing the areas where measurements were made. The statistical tests discussed in this chapter assume that the data constitute a random sample from the population. If a sample of measurements is to represent the entire site, every sampling point within the area represented by the sample must have a non-zero probability of being selected as part of the sample. If all points have an equal opportunity for selection, the sampling procedure will generate a simple random sample. Most procedures presented in this chapter require a simple random sample. Stratification of the site will usually result in differing probabilities of selection within each stratum. In this case, the sample is not a simple random sample, and a statistician should be consulted before conducting the analysis.

Judgmental (or "authoritative"[2]) samples are samples that are collected in areas suspected to have higher contaminant concentrations due to operational or historical knowledge. Judgmental samples cannot be extrapolated to represent the entire site. In some cases, there is a great deal of bias associated with the collection of judgmental samples. The statistical procedures recommended in this chapter are based on random samples and should not be used on judgmental samples. Graphical methods such as posting plots may be used to display judgmental data. These displays may reveal likely sources and pathways of contamination. Kriging[3] and other spatial smoothing algorithms may be applied to identify areas with suspected high concentrations for conducting the remediation, although the estimated mean concentrations must be recognized for their upward bias.

Depending upon the data and other site-specific considerations, statistical analysis should involve one or a combination of the following methods:

- Parametric statistical comparison methods involving comparison of one or more parameters of the distribution of site samples with the corresponding parameter of the background distribution, such as the Student t-test; or
- Nonparametric tests, such as Wilcoxon Rank Sum (WRS) test.

Method	Application	Comments
Descriptive Summary ▸ Mean ▸ Median ▸ Standard deviation ▸ Variance ▸ Percentiles ▸ Kurtosis	Preliminary examination of data for comparison with site history and land use activities in the establishment of background. Use as a preliminary screening tool.	Simple and straightforward; less statistical rigor.
Simple Comparisons	Used with very small data sets.	Not recommended
Parametric Tests ▸ Student t-test ▸ ANOVA ▸ Student t-test ▸ Behrens-Fisher Student t-test	Data must be normal or transformable to normal. Use when more data points are available (n > 10). Examine data for normality or lognormality in distribution.	Statistically robust and used frequently in parametric data analysis.
Nonparametric Tests ▸ Wilcoxon Rank Sum Test (also called the "Mann-Whitney Test") ▸ Gehan Test	Use when data are not normally distributed, as rank-ordered tests make no assumption on distribution.	Statistically robust and used frequently in background estimation.

The box at the top of this page lists some of the statistical tests and applications recommended for establishing background constituent concentrations. These tests are discussed in more detail in the following sections.

5.1 Descriptive Summary Statistics

Several statistics can be used to describe data sets. These statistics may be used in many of the tests described later in this chapter. There are two important features of a data set: *central tendency* and *dispersion*.

To describe central tendency, estimators of the mean such as arithmetic mean, median, mode, and geometric mean are employed. The sample mean is an arithmetic average for simple random sampling designs; however for complex sampling designs, such as stratification, the sample mean is a weighted arithmetic average. The sample mean is influenced by extreme values (large or small) and can easily be influenced by non-detects. The sample median value falls directly in the middle of the data when the measurements are ranked in order from smallest to largest. More simply, the median is the middlemost value in the data set. The median is less affected by the presence of values recorded as being below the detection limit.

The dispersion around the central tendency is described by such items as the range, variance, sample standard deviation, and coefficient of variation. The easiest measure of dispersion is the sample range. For small samples, the range is easy to interpret and may adequately represent the spread of the data. For large samples, the range is not very informative because it only considers and is greatly influenced by extreme values. The sample variance measures the dispersion from the mean of a data set and is affected by extreme values and by a large number of non-detects. The coefficient of variation (CV) is a unitless measure that allows the comparison of dispersion across several sets of data. The CV is often used instead of the standard deviation in environmental applications because the standard deviation is often proportional to the mean. The standard deviation is affected by values below the detection limit, and some method of substituting

numerical values for these must be found.

5.2 Simple Comparison Methods

Simple comparison methods rely on descriptive summary statistics, such as comparing means or maximums. These approaches can be used with very small data sets but are highly uncertain.

5.3 Statistical Methods for Comparisons with Background

Many statistical tests and models are only appropriate for data that follow a particular distribution. Statistical tests that rely on knowledge of the form of the population distribution for the data are known as *parametric* tests, because the test is usually phrased in terms of the parameters of the distribution assumed for the data. Two of the most important distributions for tests involving environmental data are the normal distribution and the lognormal distribution. A normal distribution has only two parameters, the mean and variance. Lognormal distributions also have only two parameters, but there are several common ways to parameterize the lognormal distribution. In this chapter, use of parametric comparison methods like t-tests or ANOVA may require normalization of data by conversion to a log scale.

Tests for the distribution of the data (such as the Shapiro-Wilk test for normality) often fail if there are insufficient data, if the data contain multiple populations, or if there is a high proportion of non-detects in the sample.[4] Tests for normality lack statistical power for small sample sizes. Therefore, for small sample sizes, nonparametric tests should be used to avoid incorrectly assuming the data are normally distributed when there is not enough information to test this assumption. Thus, when the distribution cannot be determined, it is more appropriate to use nonparametric tests.

Statistical tests that do not assume a specific mathematical form for the population distribution are called distribution-free or *nonparametric* statistical tests. Nonparametric tests have good test performance for a wide variety of distributions, and their performance is not unduly affected by outliers. Nonparametric tests can be used for normal or non-normal data sets. If one or both of the data sets fail to meet the test for normality, or if the data sets appear to come from different types of distributions, then nonparametric tests may be the only alternative for the comparison with background. However, for normal data with no outliers or non-detect values, the parametric methods discussed in the next section are somewhat more powerful. Nonparametric tests are discussed in Section 5.3.2.

The choice of a parametric test or a nonparametric test often depends on sample size. There are different circumstances that must be considered:

- If a parametric test is applied to data from a non-normal population and the sample size is large, the parametric test will work well. The central limit theorem ensures that parametric tests will work because parametric tests are robust to deviations from normal distributions as long as the sample size is large. Unfortunately, it is impossible to say how large is large enough because it depends on the nature of the particular distribution. Unless the population distribution is very peculiar, you can safely choose a parametric test when there are at least 24 data points in each group.

- If a nonparametric test is applied to data from a normal population and the sample size is large, the nonparametric test will work well. In this case, the p values tend to be a little too large, but the discrepancy is small. In other words, nonparametric tests are only slightly less powerful than parametric tests with large samples.

- If a parametric test is applied to data from a non-normal population and the sample size is small, the p value may be inaccurate because the central limit theory does not apply in this case.

- If a nonparametric test is applied to data from a non-normal population and the sample size is

small, the p values tend to be too high. In other words, nonparametric tests may lack statistical power with small samples.

In conclusion, large data sets do not present any problem. In this case the nonparametric tests are powerful and the parametric tests are robust. However, small data sets are challenging. In this case the nonparametric tests are not powerful, and the parametric tests are not robust.

5.3.1 Parametric Tests

Parametric statistical tests assume the data have a known distributional form. For example, the widely used t-test assumes a normal distribution for the data. They may also assume that the data are statistically independent or that there are no spatial trends in the data. Parametric statistical comparison methods, in the context of this guidance, involve comparison of one or more distribution parameters of site samples with corresponding parameters of the background distribution.

Tests for the distribution of the data offer clues on metals detected frequently at higher concentrations. For example, as a general rule, naturally occurring aluminum, iron, calcium, and magnesium tend to be normally distributed, while trace metals tend to have lognormal distributions.

Tests of Means

The most common method for background comparisons involves a comparison between means using t-tests or similar parametric methods. If the estimated means do not differ by a statistically significant amount (given a predetermined level of significance such as 0.05), then there is no substantial difference in the mean of the site data as compared to the mean of the background data.

To conduct a t-test, a null hypothesis must first be developed. (See Section 3.1 for developing null hypotheses.) The t-statistic calculated from the data is then compared to a critical value for the test which depends on the level of confidence selected to determine whether or not the null hypothesis should be rejected. Although the t-test is derived based on normality, the conclusion that the data do not follow a normal distribution does not discount the t-test. Generally, the t-test is robust and therefore not sensitive to small deviations from the assumptions of normality.

Any t-test should be discussed with a statistician prior to use since there are a number of variations and assumptions that can apply. The Student t-test has good application when comparing background sites to potentially contaminated sites.[5]

Methods such as Cochran's Approximation to the Behrens-Fischer Student t-test are also recommended. This statistical comparison method requires that two or more discrete samples be taken at each sampling station. Note that the choice of a specific t-test depends on site-specific information and other statistical considerations.

Tests of Outliers

The are many parametric tests for outliers, based on deviations from the normal distribut EPA's QA/G-9, *Guidance for Data Quality Assessment*[2] ion. Three

Parametric Tests		
Test	**Purpose**	**Assumptions**
t-test	Test for difference in means	Normality, equal variances
Upper Tolerance Limit (UTL)	Test for outliers	Normality
Extreme Value (Dixon's) Test	Test for one outlier	Normality, not including outlier
Rosner's Test	Test for up to 10 outliers	Normality, sample size 25 or larger
Discordance Test	Test for one outlier	Normality, not including the outlier

of these tests are explained in detail in, including Dixon's test, Rosner's test, and the Discordance test shown in the box on the previous page. In addition to these tests, suspected outliers may be identified using a tolerance limit approach. There are parametric and nonparametric forms of tolerance intervals. This section discusses the parametric version.[6] A nonparametric version of tolerance intervals is presented in Section 5.3.2.

Confidence intervals provide interval estimates for unknown population parameters. Tolerance limits differ from confidence intervals. Tolerance limits provide an interval within which at least a certain proportion of the population lies with a specified probability that the stated interval does indeed "contain" the proportion of the population. An example would be a situation in which you are trying to draw a random sample, and want to know how large the sample size should be so that you can be 95 percent sure that at least 95 percent of the population lies between the smallest and the largest observation in the sample. The tolerance limit described here would be a two-sided tolerance limit. Similarly one-sided tolerance limits can be developed. In fact, one-sided tolerance limits are identical with one-sided confidence intervals for quantiles (percentiles).

Establishing a tolerance limit (TL) is recommended for identifying outliers. A TL is a confidence limit on a percentile of the data, rather than a confidence limit on the mean. For example, a 95 percent TL for 95 percent coverage represents the value below which 95 percent of the population are expected to fall (with 95 percent confidence). *In using a TL for background comparisons, a site sample is considered to be contaminated when its concentration exceeds the upper TL of the background data set.* TL tests are fairly sensitive and require a minimum of 8 to 16 background data points. A one-sided upper TL is estimated using the mean plus a standard deviation times the tolerance coefficient (K) at the 95 percent probability level for a 95 percent coverage.

5.3.2 Nonparametric Tests

The statistical tests discussed in the previous section rely on the mathematical properties of the population distribution (normal or lognormal) selected for the comparison with background. Assumptions concerning the population distribution are difficult to verify or difficult to satisfy for both populations. Use of parametric statistical tests when the data do not follow the assumed distribution may lead to inaccurate comparisons that are adversely affected by outliers and by assumptions made for handling non-detect values.

Tests that do not assume a specific mathematical form for the underlying distribution are called distribution-free or nonparametric statistical tests. The property of *robustness* is the main advantage of nonparametric statistical tests. Robustness means that nonparametric tests have good test performance for a wide variety of distributions, and that performance is not unduly affected by outliers.

Nonparametric tests can be used for normal or non-normal data sets. If one or both of the data sets failed to meet the test for normality, or if the data sets appear to come from different types of populations, then nonparametric tests may be the only alternative for the comparison with background. If the two data sets appear to be from the same family of distributions, use of a specific statistical test that is based on this knowledge is not necessarily required because the nonparametric tests will perform almost as well. However, for normal data with no outliers or non-detect values, the parametric methods discussed in the previous section are somewhat more powerful.

Several nonparametric test procedures are available for conducting background comparisons. Nonparametric tests compare the shape and location of the two distributions instead of a statistical parameter (such as mean and median). Nonparametric tests are currently used by some EPA Regions on a case-by-case basis. These methods have varying levels of sensitivity and data requirements and should be considered as the preferred methods whenever data are heavily censored (a high percentage of non-detect values).

Nonparametric Tests	
Test	**Assumptions**
Wilcoxon Rank Sum (WRS)	Both samples are randomly selected from respective populations and mutually independent; distributions are identical (except for possible difference in location parameter).
Gehan Test	Multiple detection limits and non-detect.

Wilcoxon Rank Sum Test for Background Comparisons

The Wilcoxon Rank Sum (WRS)[7] test is an example of a nonparametric test used for determining whether a difference exists between site and background population distributions. The WRS tests whether measurements from one population consistently tend to be larger (or smaller) than those from the other population. This test determines which distribution is higher by comparing the relative ranks of the two data sets when the data from both sources are sorted into a single list. One assumes that any difference between the background and site concentration distributions is due to a shift in the site concentrations to higher values (due to the presence of contamination in addition to background).

Two assumptions underlying this test are: 1) samples from the background and site are independent, identically distributed random samples, and 2) each measurement is independent of every other measurement, regardless of the set of samples from which it came. The test assumes also that the distributions of the two populations are identical in shape (variance), although the distributions need not be symmetric.

The WRS test has three advantages for background comparisons:

- The two data sets are not required to be from a known type of distribution. The WRS test does not assume that the data are normally or log-normally distributed, although a normal distribution approximation often is used to determine the critical value for the test for large sample sizes.

- It allows for non-detect measurements to be present in both data sets.[8] The WRS test can handle a moderate number of non-detect values in either or both data sets by treating them as ties.[9] Theoretically, the WRS test can be used with up to 40 percent or more non-detect measurements in either the background or the site data. If more than 40 percent of the data from either the background or site are non-detect values, the WRS test should not be used.[10]

- It is robust with respect to outliers because the analysis is conducted in terms of ranks of the data. This limits the influence of outliers because a given data point can be no more extreme than the first or last rank.

The WRS test may be applied to either null hypothesis in the two forms of background test discussed in Chapter 3: *indistinguishable from background* or *exceed by more than a substantial difference*. In either form of background test, the null hypothesis is assumed to be true unless the evidence in the data indicates that it should be rejected in favor of the alternative.

WRS Test Procedure for Background Test Form 1

Null Hypothesis (H_o): The mean of the site distribution is less than or equal to the mean of the background ($\Delta \leq 0$).

Alternative Hypothesis (H_A): The mean of the site distribution exceeds the mean of the background distribution ($\Delta > 0$).

Procedures for Non-Detect Values in WRS Test

If there are t non-detect values, they are considered as "ties" and are assigned the average rank for this group. Their average rank is the average of the first t integers, (t+1)/2. If more than one detection limit was in use, all observations below the largest detection limit should be treated as non-detect values. Alternatively, the Gehan test may be performed.

The WRS test for Background Test Form 1 is applied as outlined in the following steps. The lead-contaminated storage yard example from Chapter 3 serves to illustrate the procedure. (Although the study team selected to use Background Test Form 2 in this example problem, both forms of the test will be evaluated.)

Hypothetical data for the storage yard example is shown in Exhibits 5.1 and 5.2 for the on-site and background areas, respectively. There is one non-detect measurement (ND) in the data collected on site

Data (mg/kg)	Source
ND	Site
34.0	Site
39.5	Site
48.6	Site
54.9	Site
70.9	Site
72.1	Site
81.3	Site
83.2	Site
86.2	Site
88.2	Site
96.1	Site
98.3	Site
104.3	Site
105.6	Site
129.0	Site
139.3	Site
156.9	Site
167.9	Site
208.4	Site

Exhibit 5.1 Site data.

Data (mg/kg)	Source	Data+50 (mg/kg)	Data+100 (mg/kg)	Source
ND	Background	50.0	100.0	Background+S
ND	Background	50.0	100.0	Background+S
ND	Background	50.0	100.0	Background+S
ND	Background	50.0	100.0	Background+S
ND	Background	50.0	100.0	Background+S
0.1	Background	50.1	100.1	Background+S
15.7	Background	65.7	115.7	Background+S
46.1	Background	96.1	146.1	Background+S
48.1	Background	98.1	148.1	Background+S
49.3	Background	99.3	149.3	Background+S
53.5	Background	103.5	153.5	Background+S
58.0	Background	108.0	158.0	Background+S
59.7	Background	109.7	159.7	Background+S
68.0	Background	118.0	168.0	Background+S
88.5	Background	138.5	188.5	Background+S
96.5	Background	146.5	196.5	Background+S
115.8	Background	165.8	215.8	Background+S
122.9	Background	172.9	222.9	Background+S
126.8	Background	176.8	226.8	Background+S
147.5	Background	197.5	247.5	Background+S

Exhibit 5.2 Background data.

and five in the background data set. The background non-detects were treated as 0 values when adding S to the background measurements. This is a more conservative approach than using ½ the detection limit or other surrogate or random numbers for the non-detect values. The WRS test is very robust to this small modification as it is unlikely that any reasonable surrogate value will affect significantly the assigned rank of the non-detects in the combined data set.

Exhibit 5.3 demonstrates the WRS test procedure for Background Test Form 1, testing the null hypothesis that the site is indistinguishable from background. The background measurements (m = 20) and the site measurements (n = 20) are ranked in a single list in order of increasing size from 1 to N, where N = m + n = 40. At the top of the list, all six non-detect values are considered as ties and are assigned an average rank of 3.5 = (6+1)/2. (See box below). The ranks for each area are shown in the two columns at the right of the exhibit. The sum of the ranks of the site measurements (W_s = 491.5) and the sum of the ranks of the background measurements (W_b = 328.5) are shown at the bottom of the exhibit. Since the sum of the first N = 40 integers is N(N+1)/2 = 40(40+1)/2 = 820, the sum of W_s plus W_b should equal this number. The sum of the ranks of the site measurements (W_s = 491.5) is the test statistic used for Background Test Form 1. The sum of the site ranks is used as the test statistic for background test from 1 because we are looking for evidence that the site distribution exceeds the background distribution. To conduct the test, W_s is compared with the critical value for the WRS test for the appropriate values of n, m, and α.[11] If W_s is greater than the tabulated critical value for the test, the null hypothesis that the site is indistinguishable from background is to be rejected.

Exhibit 5.6 shows the critical values for the WRS test for selected values of α for data sets with n = m = 20. The critical value for α = 10 percent is 458, and the critical value for α = 5 percent is 471. Since W_s exceeds the critical values for most commonly used values of α, the null hypothesis is rejected. Hence, the site is distinguishable from background at a confidence level of 95 percent. Note that the null hypothesis would not be rejected at α = 1 percent.

WRS Test Procedure for Background Test Form 2

Null Hypothesis (H_o): The site distribution exceeds the background distribution by more than a substantial difference S ($\Delta > S$).

Alternative Hypothesis (H_A): The site distribution does not exceed the background distribution by more than S ($\Delta \leq S$).

The WRS test for Background Test Form 2 is applied as outlined in the following steps. The lead example will again serve as an illustration of the procedure. In the example from Chapter 3, the study team chose to use Background Test Form 2, with α = 10 percent and a substantial difference of S = 100 mg/kg. First, the background measurements are adjusted by adding S = 100 mg/kg to each measured value. Exhibit 5.2 contains two columns on the right which show the S-adjusted background data for S = 50 mg/kg and S = 100 mg/kg.

The adjusted background measurements and the measurements from the site are ranked in increasing order from 1 to 40 in Exhibit 5.4. Note that the five adjusted background measurements that were non-detects are tied at 100 mg/kg. They are all assigned the average rank of 16 for that group of tied measurements.

The sum of the ranks of the adjusted measurements from background, W_b = 544, is the test statistic for Background Test Form 2. Note that the test statistic for Background Test Form 2 differs from the test statistic for Background Test Form 1. In this case, we are looking for evidence that S plus the background distribution is greater than the site distribution. Earlier, in Background Test Form 1, we were looking for evidence that the site distribution exceeds the (unmodified) background distribution. The critical value for the WRS test in Exhibit 5.6 for α = 10 percent is 458. Since W_b is greater than the critical value, the null hypothesis that the site exceeds background by more than a substantial difference of 100 mg/kg is rejected at the 90 percent

Rank	Data (mg/kg)	Source	Ranks for Site	Ranks for Background
3.5	ND	Site	3.5	
3.5	ND	Background		3.5
3.5	ND	Background		3.5
3.5	ND	Background		3.5
3.5	ND	Background		3.5
3.5	ND	Background		3.5
7	0.1	Background		7
8	15.7	Background		8
9	34.0	Site	9	
10	39.5	Site	10	
11	46.1	Background		11
12	48.1	Background		12
13	48.6	Site	13	
14	49.3	Background		14
15	53.5	Background		15
16	54.9	Site	16	
17	58.0	Background		17
18	59.7	Background		18
19	68.0	Background		19
20	70.9	Site	20	
21	72.1	Site	21	
22	81.3	Site	22	
23	83.2	Site	23	
24	86.2	Site	24	
25	88.2	Site	25	
26	88.5	Background		26
27	96.1	Site	27	
28	96.5	Background		28
29	98.3	Site	29	
30	104.3	Site	30	
31	105.6	Site	31	
32	115.8	Background		32
33	122.9	Background		33
34	126.8	Background		34
35	129.0	Site	35	
36	139.3	Site	36	
37	147.5	Background		37
38	156.9	Site	38	
39	167.9	Site	39	
40	208.4	Site	40	
820		Sum of Ranks	491.5	328.5
			W_s	W_b

Exhibit 5.3 WRS test for Test Form 1
H_0: site < background

Rank	Data (mg/kg)	Source	Ranks for Site	Ranks for Background + 100
1	ND	Site	1	
2	34.0	Site	2	
3	39.5	Site	3	
4	48.6	Site	4	
5	54.9	Site	5	
6	70.9	Site	6	
7	72.1	Site	7	
8	81.3	Site	8	
9	83.2	Site	9	
10	86.2	Site	10	
11	88.2	Site	11	
12	96.1	Site	12	
13	98.3	Site	13	
16	100.0	Background+S		16
16	100.0	Background+S		16
16	100.0	Background+S		16
16	100.0	Background+S		16
16	100.0	Background+S		16
19	100.1	Background+S		19
20	104.3	Site	20	
21	105.6	Site	21	
22	115.7	Background+S		22
23	129.0	Site	23	
24	139.3	Site	24	
25	146.1	Background+S		25
26	148.1	Background+S		26
27	149.3	Background+S		27
28	153.5	Background+S		28
29	156.9	Site	29	
30	158.0	Background+S		30
31	159.7	Background+S		31
32	167.9	Site	32	
33	168.0	Background+S		33
34	188.5	Background+S		34
35	196.5	Background+S		35
36	208.4	Site	36	
37	215.8	Background+S		37
38	222.9	Background+S		38
39	226.8	Background+S		39
40	247.5	Background+S		40
820		Sum of Ranks	276	544
			W_s	W_b

Exhibit 5.4 WRS test for Test Form 2
H_0: site > background + 100

confidence level.

Exhibit 5.5 shows the WRS test for the lead example using Background Test Form 2 with a smaller (more conservative) value for a substantial difference, S = 50 mg/kg. The sum of the ranks of the S-adjusted background measurements is W_b = 441. After examination of Exhibit 5.6, it is clear that the null hypothesis that the site exceeds background by more than 50 mg/kg cannot be rejected at any reasonable level of confidence.

Rank	Data (mg/kg)	Source	Ranks for Site	Ranks for Background + 50
1	ND	Site	1	
2	34.0	Site	2	
3	39.5	Site	3	
4	48.6	Site	4	
7	50.0	Background+S		7
7	50.0	Background+S		7
7	50.0	Background+S		7
7	50.0	Background+S		7
7	50.0	Background+S		7
10	50.1	Background+S		10
11	54.9	Site	11	
12	65.7	Background+S		12
13	70.9	Site	13	
14	72.1	Site	14	
15	81.3	Site	15	
16	83.2	Site	16	
17	86.2	Site	17	
18	88.2	Site	18	
19	96.1	Background+S		19
20	96.1	Site	20	
21	98.1	Background+S		21
22	98.3	Site	22	
23	99.3	Background+S		23
24	103.5	Background+S		24
25	104.3	Site	25	
26	105.6	Site	26	
27	108.0	Background+S		27
28	109.7	Background+S		28
29	118.0	Background+S		29
30	129.0	Site	30	
31	138.5	Background+S		31
32	139.3	Site	32	
33	146.5	Background+S		33
34	156.9	Site	34	
35	165.8	Background+S		35
36	167.9	Site	36	
37	172.9	Background+S		37
38	176.8	Background+S		38
39	197.5	Background+S		39
40	208.4	Site	40	
820		Sum of Ranks	379	441
			W_s	W_b

Exhibit 5.5 WRS test for Test Form 2
H_0: site > background + 50

In conclusion, site concentrations in this example are significantly higher than background concentrations. The site distribution may exceed background by 50 mg/kg or more, but it is unlikely that the site distribution is more than 100 mg/kg above background.

α	Critical Value
0.20	442
0.15	449
0.10	458
0.05	471
0.025	482
0.010	495
0.005	504
0.001	521

Exhibit 5.6 Critical Values for the WRS Test for n = m = 20

Gehan's Form of the WRS Test

The Gehan test is a generalized version of the WRS test.[12] If there are a large number of non-detect measurements and several different detection levels, Gehan's form of the WRS test is a more powerful test for the background comparison. The Gehan test addresses multiple detection limits using a modified ranking procedure rather than relying on the "all ties get the same rank" approach used in the WRS test. After the modified ranking is completed, the standard WRS test procedure discussed above is applied to determine if the null hypothesis should be rejected. It has been recommended that there should be at least 10 data values in each data set to use this test.

Walsh's Tests for Outliers

Nonparametric tests for outliers are summarized in the box above. Walsh's test is a nonparametric test for determining the presence of outliers in either the background or on-site data sets. This test was developed to detect up to a specified number of outliers, r. The test requires large sample sizes (n > 60 for α = 10 percent; and n > 220 for α = 5 percent). Procedures for conducting this test is discussed in Section 4.4 of QA/G-9, *Guidance for Data Quality Assessment*.[2]

Nonparametric Tolerance Intervals

The parametric tolerance intervals discussed in

Section 5.2 are derived based on the assumption of a normal distribution. If the data are not normal and are not easily transformed to normality, then non-parametric tolerance intervals may be calculated for the background distribution to provide a tolerance level for screening site data. The use of nonparametric tolerance intervals is explained as follows:

By definition, a single random sample from any distribution has a probability of 0.1 of exceeding the 90th percentile of the distribution (x.90). The probability that the maximum of two independent random samples from the same distribution will exceed the 90th percentile is calculated using probability theory. The probability that the maximum exceeds the 90th percentile equals 1 minus the probability that both samples are less than the 90th percentile:

$$\Pr\{ \text{Max}(X_1, X_2) > x.90 \} = 1 - \Pr\{ X_1 < x.90$$

and

$$X_2 < x.90 \} = 1 - (.90)(.90) = 1 - (.90)^2 = 0.19$$

In general we have,

$$\Pr\{ \text{Max}(X_1, \ldots , X_n) > x.90 \} = 1 - (.90)^n$$

This probability increases as the number of samples increases. In a sample size of 22, there is more than a 90 percent chance that the maximum in the sample will exceed the 90th percentile of the underlying distribution. In this case,

$$\Pr\{ \text{Max}(X_1, \ldots , X22) > x.90 \} = 0.9015$$

If we take 22 independent samples, the maximum of these samples will usually exceed the 90th percentile. So if the maximum of 22 samples is less than the level of concern, then there is at least a 90 percent probability that the 90th percentile of the population distribution is also less than the level of concern. In other words, there is less than a 10 percent chance that the maximum of 22 samples would fail to exceed the 90th percentile of the distribution, and this is the only way 90th percentile could exceed the level of concern if the maximum does not. Therefore, the maximum value in a sample size of 22 is said to provide a one-sided 90 percent non-parametric tolerance interval for the 90th percentile of the population distribution. Similarly, the maximum value in a sample size of 45 provides a 90 percent nonparametric tolerance interval for the 95th percentile of the population. Similarly, n = 58 gives a 95 percent TL for the 95th percentile.

5.4 Hypothesis Testing

Hypothesis testing was discussed in detail in Section 3. Here, some of this information is reviewed, and additional aspects of such testing are discussed.

5.4.1 Initial Considerations

For Superfund sites, use of a null hypothesis and alternative hypothesis is recommended when comparing data sets from contaminated areas with background data. For example, a null hypothesis could be “there is no difference between the mean contaminant concentration in samples from contaminated areas and background data sets.” The alternative hypothesis would be “there is a difference between mean contaminant concentration in samples from contaminated areas and background data sets.” To conduct the comparison, parametric or nonparametric statistical tests are recommended. Use of parametric comparison methods like t-tests or ANOVA may require normalization of data such as the conversion to a log scale. Depending upon the data and other site-specific considerations, statistical analysis should involve one or a combination of the following methods:

- A preliminary descriptive analysis involving the comparison of median, mean, and upper range concentrations between sample sets considered site-related and background;
- Parametric statistical comparison methods involving the comparison of one or more parameters of the distribution of site samples with corresponding parameters of the (assumed or

sampled) background distribution, such as Gosset's Student t-test or Cochran's Approximation to the Behrens-Fischer Student t-test; or

- Nonparametric tests, such as the Wilcoxon Rank Sum test (on a case-by-case basis).

Once a test has been selected, the assessor must consider several questions:

- *What should the null and alternative hypotheses be?* What are we testing? What are we trying to prove or disprove about the site and background?

- *Should the test be one-tailed or two-tailed?* Should we ask whether the site and background are from the same population, or should we focus on whether one is more contaminated than the other?

- *What confidence level should be used?* At what "cut-off" point do we accept or reject the hypothesis?

5.4.2 Examples

It may be easiest to explore these questions by using an example. Suppose we have an area that meets our criteria for local background (unaffected by site operations). The data from this area for Chemical X (mg/kg) are as follows:

66 67 68 68 69 69 69
70 70 70 71 71 71 72
72 72 72 73 74 74 75

These data were collected randomly and are normally distributed. There are 21 measurements (n = 21), with an average of 70.6 mg/kg and a standard deviation of 2.37 mg/kg.

We also have data from an on-site process area. These data for Chemical X (mg/kg) are as follows:

62 63 64 65 66 68 68
69 69 70 71 71 72 72
72 73 74 75 77 78 80

These data were collected randomly and are normally distributed. There are 21 measurements (n = 21), with an average of 70.4 mg/kg and a standard deviation of 4.86 mg/kg.

We can see that the background and on-site areas appear to be similar, but some of the on-site data exceed the background data. We would like to be able to state with a given level of confidence whether the data are essentially from the same population, or not. If we use the t test to compare the true means of these data sets, we could test the hypothesis that the background mean and the site mean are essentially equal (H_0, the null hypothesis). If H_0 is not true, then we would support the alternative hypothesis that the means are not equal. This is a two-tailed test, because H_0 could be rejected if the site mean is greater than the background mean or if the site mean is less than background mean.

Example 1: $H_0: \mu_s = \mu_b$
$H_A: \mu_s \neq \mu_b$

(Note that this is a two-tailed version of Test Form 1.) Using the equations in EPA's QA/G-9, *Guidance for Data Quality Assessment*[13], for t, we find that t = 0.1693. At 40 degrees of freedom[14], for a two-tailed test, our t falls below the t of 0.681, where $\alpha = 0.5$. Therefore, if we had chosen an α of 0.01 (99 percent confidence), 0.05 (95 percent confidence), or 0.1 (90 percent confidence), we would not reject our null hypothesis. Only if we were testing at less than 50 percent confidence would we reject H_0.

Although the t-test is specifically designed for the normal distributions used in this simplified example, other tests also may be used to compare these data sets. If we consider a nonparametric comparison of these data using a one-tailed version of Test Form 1,

Example 2: $H_0: \mu_s \leq \mu_b$
$H_A: \mu_s > \mu_b$,

the WRS test statistics are $W_s = 445$ and $W_b = 458$ with n = m = 21. The nearly equal sums of ranks indicates that the two distributions have similar

locations, and H_0 cannot be rejected in favor of HA. In this example, the site mean is only slightly smaller than the background mean, so the inconclusive test results are not unexpected.

In other situations, the site mean may exceed the background mean although the test indicates that there is no evidence for rejecting the hypothesis that the site is clean. When using Background Test Form 1, the confidence level of the test determines the rate of Type I decision errors (false rejection of the null hypothesis) while the sample size determines the rate of Type II decision errors (false acceptance of the null hypothesis). Hence, although a level-α test may indicate that there is no statistical evidence for rejecting the null hypothesis that the site is clean, it is important to examine the retrospective power of the test to determine if the data sets have sufficient power to reject the null hypothesis when it is false. If the power is inadequate, then reports of the test results should indicate that the data had insufficient power to determine if the site exceeds background or not. When using Test Form 1, the higher the confidence limit, the more likely this test is to find that the site is clean (from the same population as background). Choosing the rejection range for the hypothesis involves balancing both kinds of error.

In general, EPA recommends a minimum confidence limit of 80 percent and a maximum confidence limit of 95 percent.

Suppose we want to compare our background data set with another on-site process area. The data for Chemical X (mg/kg) are as follows:

56 58 60 62 66 67 68
70 72 73 75 76 81 82
84 85 87 90 91 92 103

These data were collected randomly and are normally distributed. There are 21 measurements (n = 21), with an average of 76.1 mg/kg and a standard deviation of 12.68 mg/kg.

Is this area significantly different from background? The arithmetic mean is 76.1 mg/kg, compared to the background mean of 70.6 mg/kg. But is this difference truly significant? After all, the mean of the first process area, 70.4 mg/kg, was different from the background mean. According to the t test, however, we did not find the difference of 0.2 mg/kg to be significant at the 80-99 percent confidence levels. What about the second process area?

Suppose we decide that what we are really interested in is whether the site is dirty (above background). Instead of a 2-tailed test, we could perform a 1-tailed test:

Example 3: $H_0: \mu_s > \mu_b$
$H_A: \mu_s \leq \mu_b$

(Note that this is Test Form 2 with S = 0.) This test is 1-tailed because the rejection region is only on one side of the distribution; that is, we are only interested in whether the site is greater than the background.

For comparison, the nonparametric WRS test is applied by ranking the two data sets in a single list, then summing the ranks for the site and background measurements separately. The WRS test statistics for this test are $W_b = 395.5$ and $W_s = 507.5$. Note that $W_b + W_s = 903 = (n+m)(n+m+1)/2$. The higher sum of ranks for the site measurements indicates that the site distribution exceeds the background distribution, indicating a very small likelihood of rejecting the null hypothesis. To conduct a WRS test using this pair of hypotheses, the sum of the background ranks (W_b) is compared to the critical value for the test, and the null hypothesis is rejected if $W_b > W_{crit}$. The test requires that the background ranks be significantly higher than those for the onsite data. The critical values for the WRS test are 502.4 at the 90 percent confidence level and 516.9 at 95 percent confidence for n = m = 21[15].

To use the normal distribution theory correctly, for a 1-tailed t test, with 40 degrees of freedom, the t of -1.95 is calculated for the background mean minus the site mean. This t falls between the 95 percent and 97.5 percent confidence levels. If we were

testing at 80 percent or 95 percent confidence, we would reject H_0 and find that the site is less than or equal to background—in other words, "clean." At 99 percent confidence, H_0 could not be rejected. In this case, therefore, a lower confidence limit seems to *increase* the chances of finding that the site is clean, where in our earlier tests we found that a lower confidence limit *decreased* the chances of considering the site clean. Why is this?

The difference is in the setup of the hypotheses. In the first case (examples 1 and 2), the null hypothesis was that the site and background were from the same population (the site was clean). In the later case, the null hypothesis was that the site mean exceeded the background mean (the site is dirty). In essence, we have shifted the burden of proof. If we are really interested in whether the site is dirty (greater than background), then our last test could have looked at these hypotheses:

Example 4: $H_0: \mu_s \leq \mu_b$
$H_A: \mu_s > \mu_b$

(Note that this is a one-tailed version of Test Form 1.) Using the site mean minus the background mean for this test, we derive a t of 1.95. At the 80 percent confidence level, we would reject H_0 and find that the site is dirty. At the 95 percent confidence level and above, we would accept H_0 and find that the site is clean because the data are insufficient to support this higher level of confidence demanded for rejection. Once again, with Test Form 1, a *lower* confidence level results in a *more conservative* approach to environmental protection.

There is another problem, besides burden-of-proof, with Example 3. As discussed in Chapter 3, the null hypothesis that there *is* a substantial difference (Test Form 2, $\Delta > 0$) should only be tested if some minimal difference (S) is specified. This is because the null hypothesis $H_0: \Delta > 0$ (i.e., $H_0: \mu_s > \mu_b$) will be rejected only if the site mean is significantly below the background mean. In a more typical case, the site mean may be almost equal to or slightly below the background mean, and the null hypothesis will only be rejected when a large number of samples is collected to reduce the uncertainty to below the magnitude of the difference in means.

Using a one-tailed Test Form 2, we can test whether the site exceeds background by more than S using the hypotheses:

Example 5: $H_0: \mu_s \geq \mu_b + S$
$H_A: \mu_s < \mu_b + S$

Using a substantial difference of S = 12 (approximately 1/6 of background) for the t-test of these hypotheses, we obtain the WRS test statistics $W_{b+S} = 521$ and $W_s = 382$, indicating at first glance that the site distribution is considerably lower than the background distribution plus 12. As noted above, the critical values for the WRS test are 502.4 at the 90 percent confidence level and 516.9 at 95 percent. Therefore, the null hypothesis that the site exceeds background by more than S is rejected at either level of confidence. For this test, the confidence level should be at least 80 percent; for a more conservative test, use higher levels of the confidence range.

5.4.3 Conclusions

Now we return to our original three questions. Exhibit 5.7 also summarizes this information.

- *What should the null and alternative hypotheses be?*
- *Should the test be one-tailed or two-tailed?*
- *What confidence level should be used?*

To determine whether the site and background are from the same population, these hypotheses can be used in a two-tailed Test Form 1:

$H_0: \mu_s = \mu_b$
$H_A: \mu_s \neq \mu_b$

For this test, the confidence level should be at least 80 percent but no more than 95 percent. For a more conservative test, use the lower end of the confidence range.

To determine whether the site is significantly

greater than background, these hypotheses can be used in a one-tailed Test Form 1:

$$H_0: \mu_s \leq \mu_b$$
$$H_A: \mu_s > \mu_b$$

For this test, the confidence level should be at least 80 percent; for a more conservative test, use the lower end of the confidence range and require adequate power.

If testing the hypotheses in reverse—Test Form 2—to show whether the site is greater than background + S, use a higher confidence level, such as 95 percent, and specify a substantial difference S. (See Appendix for guidance on choosing S.) To determine whether the site exceeds background by more than S, these hypotheses can be used in a one-tailed Test Form 2:

$$H_0: \mu_s \leq \mu_b + S$$
$$H_A: \mu_s < \mu_b + S$$

For this test, the confidence level should be at least 80 percent; for a more conservative test, use higher levels of the confidence range.

What to test:	H_0	H_A	Recommended alpha	Rejection criteria
H_0: site and background are from the same population; *vs.* H_A: site and background are from different populations (Two-tailed, Test Form 1)	$\mu_s = \mu_b$	$\mu_s \neq \mu_b$	80-95% confidence (α = 0.2 to 0.05) [More conservative: α = 0.2]	For 2-sided t test, e.g., reject H_0 if $t > t_{\alpha/2}$ or if $t < -t_{\alpha/2}$
H_0: site is less than or from the same population as background; *vs.* H_A: site is greater than background (One-tailed, Test Form 1)	$\mu_s \leq \mu_b$	$\mu_s > \mu_b$	80-95% confidence (α = 0.2 to 0.05) [More conservative: α = 0.2]	For 1-sided t test, e.g., reject H_0 if $\lvert t \rvert > t_\alpha$ For 1-sided t test, e.g., reject H_0 if $t > t_\alpha$*
H_0: site is greater than background + S; *vs.* H_A: site is less than or from the same population as background + S (One-tailed, Test Form 2)	$\mu_s \leq \mu_b+S$	$\mu_s < \mu_b+S$	80-95% confidence (α = 0.2 to 0.05) [More conservative: α = 0.05]	For 1-sided t test, e.g., reject H_0 if $\lvert t \rvert > t_\alpha$ For 1-sided t test, e.g., reject H_0 if $t < -t_\alpha$*

* Assuming the test statistic, t, is calculated using site mean minus background mean (or background mean + S, for Test Form 2) in the numerator

Exhibit 5.7 What to test.

CHAPTER NOTES

1. Gilbert, R.O. & J.C. Simpson. June 1994. *Statistical Methods for Evaluating the Attainment of Cleanup Standards, Volume 3*. EPA 230-R-94-004.

2. U.S. Environmental Protection Agency (EPA). July 2000. *Guidance for Data Quality Assessment: Practical Methods for Data Analysis, EPA QA/G-9, QA00 Version*. Quality Assurance Management Staff, Washington, DC, EPA 600-R-96-084. Available at http://www.epa.gov/quality/qa_docs.html. See Section 1.3.1 for guidance on "authoritative samples."

3. Cressie, N. 1991. *Statistics for Spatial Data*, New York: John Wiley & Sons. See Section 3.2.

4. When using parametric statistical tests, a limit of 15 percent non-detect measurements in either data set is suggested in the Navy's *Procedural Guidance for Statistically Analyzing Environmental Background Data*. Nonparametric statistical methods are recommended if this limit is exceeded.

5. Michigan Department of Environmental Quality Waste Management Division. April 1994. *Guidance Document: Verification of Soil Remediation*. Revision 1. http://www.deq.state.mi.us/wmd/docs/vsr.html.

6. J.L. Devore, 2000. *Probability and Statistics for Engineering and the Sciences*, 5th Ed., Duxbury Press, Pacific Grove, California.

7. The WRS test is also called the Mann-Whitney test, which is mathematically equivalent to the WRS test. Sometimes, the combined name is used: Wilcoxon-Mann-Whitney test.

8. In general, the use of "non-detect" values in data reporting is not recommended. Wherever possible, the actual result of a measurement, together with its uncertainty, should be reported. Estimated concentrations should be reported for data below the detection limit, even if these estimates are negative, because their relative magnitude compared to the rest of the data is of importance.

9. The Gehan test discussed in the next section should be considered if there are many non-detect values with different detection levels.

10. A limit of 50 percent non-detect values is suggested in the Navy's *Procedural Guidance for Statistically Analyzing Environmental Background Data.* A more conservative limit of 40 percent non-detect values is recommended in the *Multi-Agency Radiation Survey and Site Investigation Manual (MARSSIM).*

11. Critical values for the WRS test are available in many published texts and reference books. Two sources are *Practical Nonparametric Statistics, 2nd Ed.*, W.J. Conover 1980. John Wiley & Sons, New York; and *CRC Standard Probability and Statistics Tables and Formulae*, D. Zwillinger and S. Kokoska. 2000. Chapman and Hall/CRC Press, Boca Raton, Florida.

12. See, for example: Millard, W.P., and S.J. Deverel. 1988. "Non-Parametric Statistical Methods for Comparing Two Sites Based on Data with Multiple Non-Detect Limits." *Water Resources Research*, 24:12, p 2087-2098.

13. U.S. EPA, 2000, *Guidance for Data Quality Assessment: Practical Methods for Data Analysis, Op. Cit.*, Section 3.3.1.1.

14. In this context, *degrees of freedom* (n - 1) is the number of independent observations ("n") minus the number of independent parameters estimated in computing the variation. The shape of the t-distribution curve depends upon the number of degrees of freedom. Distributions with fewer degrees of freedom have heavier tails.

15. Most readily available tables for the WRS test only extend up to sample sizes of n = m = 20. Critical values for the WRS test when n and m exceed 20 may be calculated from the large sample approximation:

$$W_{crit} = m(N+1)/2 + z_{\alpha}[\ nm(N + 1)/12\ ]^{1/2}$$

where N = n + m and z_{α} is the $100(1 - \alpha)^{th}$ percentile of the standard normal distribution. The first term is the expected value of the sum of ranks W, calculated under the assumption that the null hypothesis is true. The second term is a standard normal variate times the standard deviation of W, under the same assumptions. The first factor in the expectation term m represents the number of ranks that were summed, each having expectation (N+1)/2 under the equality assumption included in the null hypothesis. *Note that the second factor in the first term "(N+1)" is misprinted as a factor of "n" in the instructions for applying the WRS test with a large sample approximation Box 3-22 of EPA QA/G-9, QA00 Version dated July, 2000.* The correct formula above is not symmetric when n and m are exchanged.

ADDENDUM

POLICY CONSIDERATIONS FOR THE APPLICATION OF BACKGROUND DATA IN RISK ASSESSMENT AND REMEDY SELECTION

[In preparation.]

APPENDIX

ISSUES REGARDING BACKGROUND COMPARISONS FOR SUPERFUND ASSESSMENTS: "S" VALUE

A.1 Precedents for Selecting a Background Test Form

When comparing the two forms of background tests, it is important to distinguish between the selection of the null hypothesis, which is a burden-of-proof issue, and the selection of an appropriate value for a "substantial difference." It is also important to distinguish between the value that characterizes a "substantial difference over background" and the appropriate risk-based "action level" for the chemical of concern.

Existing guidance in the data quality objectives (DQO) process for choosing the null hypothesis has focused on the burden-of-proof, when the contaminant concentration is to be compared to a fixed, risk-based action level, L. The choice of Test Forms for this type of decision includes either

$$\text{a) } H_0: X < L \textit{ vs. } H_A: \quad X > L$$

or

$$\text{b) } H_0: X > L \textit{ vs. } H_A: \quad X < L,$$

where X represents the parameter of interest for the distribution of contaminant concentrations in contaminated areas. Hypothesis test *a* compares the site concentrations to the action level using a null hypothesis that the site does not exceed the action level and an alternative hypothesis that the site exceeds the action level. Hypothesis test *b* is the opposite of test *a*, using a null hypothesis, the site exceeds the action level. Background issues are not addressed directly in this framework.

One way to address background comparisons is to reformulate the hypotheses using the difference (delta—Δ) between the distribution of contaminant concentrations and background:

$$\text{a') } H_0: \Delta < S \textit{ vs. } H_A: \quad \Delta > S$$

and

$$\text{b') } H_0: \Delta > S \textit{ vs. } H_A: \quad \Delta < S.$$

In hypothesis tests *a′* and *b′*, concentrations in contaminated areas and in background locations are compared to determine if there is or is not a substantial difference between the two areas. Test *a′* uses the null hypothesis that the site does not exceed background by more than a substantial difference, while the opposite test *b′* uses the null hypothesis that the site exceeds background by more than a substantial difference (S). Approaches for selecting a value for S are addressed in the following section. Note that Test Form *b′* is the one discussed in Section 3.2.2 (Background Test Form 2); see note 6 following Chapter 3.

Background Test Form 1 focuses interest on comparisons using a "substantial" difference of S = 0. In this case, the two alternative tests are

$$\text{a'') } H_0: \Delta < 0 \textit{ vs. } H_A: \quad \Delta > 0$$

and

$$\text{b'') } H_0: \Delta > 0 \textit{ vs. } H_A: \quad \Delta < 0.$$

Background Test Form 1 (Section 3.2.1) is identical with test *a"*. This discussion demonstrates that the two background tests addressed in this paper are not opposite forms of the same test in the same sense that tests *a* and *b* are opposite forms of the same test with the same threshold. Since the guidance reviewed in this section compares opposite forms of tests with the same action level, the guidance does not contain a direct recommendation for choosing between Test Forms 1 and 2.

EPA QA/G-9[1] (Section 1.2) provides the following guidance on the selection of an appropriate null hypothesis in a choice between Test Forms *a* and *b*:

> *The decision on what should constitute the null hypothesis and what should be the alternative is sometimes difficult to ascertain. In many cases, this problem does not arise because the null and alternative hypotheses are determined by specific regulation. However, when the null hypothesis is not specified by regulation, it is necessary to make this determination. The test of hypothesis procedure prescribes that the null hypothesis is only rejected in favor of the alternative, provided there is overwhelming evidence from the data that the null hypothesis is false. In other words, the null hypothesis is considered to be true unless the data show conclusively that this is not so. Therefore it is sometimes useful to choose the null and alternative hypotheses in light of the consequences of possibly making an incorrect decision between the null and alternative hypotheses. The true condition that occurs with the more severe decision error (not what would be decided in error based on the data) should be defined as the null hypothesis. For example, consider the two decision errors: "decide a company does not comply with environmental regulations when it truly does" and "decide a company does comply with environmental regulations when it truly does not." If the first decision error is considered [the] more severe decision error, then the true condition of this error, "the company does comply with the regulations" should be defined as the null hypothesis. If the second decision error is considered the more severe decision error, then the true condition of this error, "the company does not comply with the regulations" should be defined as the null hypothesis.*

For background comparisons, that guidance may be extrapolated. When deciding between Test Forms *a"* and *b"*, there are two possible decision errors:

(i) decide the site exceeds background when it truly does not; and

(ii) decide the site does not exceed background when it truly does.

Decision error (i) occurs when a "clean" site is wrongly rejected. If decision error (i) is more serious than decision error (ii), and if the choice is between tests *a"* and *b"* with a substantial difference of 0, then Background Test Form 1 (*a"*) should be selected.

When deciding between Test Forms *a'* and *b'*, there are two possible decision errors:

(i) decide the site exceeds background + S when it truly does not; and

(ii) decide the site does not exceed background + S when it truly does.

The two Background Test Forms differ both in terms of burden of proof and in the choice of a substantial difference:

- Test Form 1 uses a conservative value for a substantial difference of S = 0, but relaxes the burden of proof by selecting the null hypothesis that the contaminant concentrations on site are indistinguishable from background.
- Test Form 2 requires a stricter burden of proof, but permits a larger value for a substantial difference.

Decision error (ii) occurs when a truly contaminated site goes undetected. If decision error (ii) is

considered more serious than error (i) and the choice is between tests *a″* and *b″* with a substantial difference of S, then Background Test Form 2 should be selected. Note that this logic does not provide a direct comparison of the two forms of background tests considered here, but does indicate situations when Test Forms 1 or 2 may be recommended over their respective opposites.

Chapter 6 of EPA QA/G-4[2] is more succinct and definitive for deciding between Test Form *a* and *b*:

> ***Define the null hypothesis (baseline condition) and the alternative hypothesis and assign the terms "false positive" and "false negative" to the appropriate decision error****. In problems that concern regulatory compliance, human health, or ecological risk, the decision error that has the most adverse potential consequences should be defined as the null hypothesis (baseline condition). In statistical hypothesis testing, the data must conclusively demonstrate that the null hypothesis is false. That is, the data must provide enough information to authoritatively reject the null hypothesis (disprove the baseline condition) in favor of the alternative. Therefore, by setting the null hypothesis equal to the true state of nature that exists when the more severe decision error occurs, the decision maker guards against making the more severe decision error by placing the burden of proof on demonstrating that the most adverse consequences will <u>not</u> be likely to occur.*

This suggests that environmental concerns are not like the jury trial process, and that the "innocent until proven guilty" assumption is an environmentally risky approach. From this viewpoint, a more protective approach would be to presume guilt, and demand proof of innocence: "guilty until proven innocent." Remember that this comparison assumes that opposite forms of the same test (*a* and *b*) are being evaluated. Extrapolation of this logic to the background problem would indicate that Test Form 2 is preferred over its true opposite, but Test Form 1 is not preferred over its opposite.

EPA guidance[3] adopts a conservative approach by stating that when the results of the investigation are uncertain, erroneously concluding that the sample area does not attain the cleanup standard is preferable to concluding that the sample area attains the cleanup standard when it actually may not. Again the recommended approach favors protection of human health and the environment.

A.2 Options for Establishing the Value of a Substantial Difference

Selection of an appropriate value to represent a substantial difference when testing for differences between concentrations in contaminated areas and background areas depends on the intended application of the test and a variety of factors. These factors include site and background variability and appropriate cleanup goals.

In this document, the term "substantial difference" (*S*) is defined as the difference in mean concentration in contaminated areas over background levels that presents a "substantial risk." Alternatively, *S* may represent a selected "not-to-exceed" action level that is appropriate for the decision at hand. *S* is measured in concentration units above the mean background concentration. The decision to use a specific value for a substantial difference may be based on direct risk assessment, a generic regulatory value, or other level selected to reflect site-specific conditions.

In situations where regulatory requirements indicate that contamination at or below background concentrations presents an unacceptably high risk, it is not possible to define a reasonable level for a substantial difference. The methodologies presented in this document are not appropriate for analysis of these unusual situations.

Background comparisons may be conducted at various stages of site characterization and remediation cycle. In the characterization stage, areas with some likelihood of contamination may be compared

to background areas to determine if contamination is present in excess of background levels. For example, the goal at this stage may be to determine the areal extent of contamination on a large site. The site is divided into sub-units that are compared to background to determine if contamination is present in the sub-unit. At this stage, Background Test Form 1 is useful for determining if the difference between the site mean and the background mean is significantly greater than zero. An upper bound for the minimum detectable difference (MDD) of the test is set by determining a value of the substantial difference *S* which will represent a threshold value for identifying possibly contaminated sub-units.

Later in the site evaluation process, background comparisons may be used to determine if a sub-unit with known contamination has been sufficiently remediated. At this stage, Background Test Form 2 is useful to demonstrate that the remediation was successful. If the goal of the remediation is to reduce contamination to near-background levels, than an appropriate value of *S* is selected that will represent the maximum amount by which a remediated sub-unit may exceed background.

A.2.1 Proportion of Mean Background Concentration

One choice for selecting a value of S is to use a specified proportion of typical mean background concentrations for the contaminant of concern:

$$S = rM_b$$

where M_b is the mean background concentration and *r* is the specified proportion. This choice may be appropriate for determining if contamination exists in a sub-unit, or if a sub-unit has been remediated successfully. There is no theoretical reason for restricting *r* to proportions less than 1, if background concentrations are far below the level that presents a substantial risk. Values of *r* near 1 may require a high number of samples, because the MDD for the test should be less than S.

The required sample size is determined by MDD/σ, where σ is the standard deviation of the concentrations in contaminated areas. Even if the area has little or no contamination, then σ will be approximately as large as the background standard deviation, which is usually at least as large as the background mean. Hence, if *r* is less than 1, then it is very likely that MDD/σ also is less than 1. If there is contamination in the contaminated area, then MDD/σ will be much less than 1.

A.2.2 A Selected Percentile of the Background Distribution

Due to the high variability in background concentrations of many chemicals, defining S as a fraction of the mean background concentration may not be appropriate. Another choice for a value to represent a substantial difference is to use a specified percentile of the distribution of background concentrations for the contaminant of concern:

$$S = (B_p - M_b)$$

where B_p is the p^{th} percentile of the background distribution and M_b is the mean background concentration. Values of p less than 0.85 may require a high number of samples, because the MDD for the test should be less than *S*. This is because the 85^{th} percentile is approximately 1 standard deviation above the background mean. When there is little or no contamination in the contaminated area, *S* is approximately equal to σ, and hence, MDD/σ usually will be near 1. If there is contamination in the contaminated area, then MDD/σ will be much less than 1.

A.2.3 Proportion of Background Variability

A third choice for selecting a value to represent a substantial difference is to use a specified proportion of variance of background concentrations for the contaminant of concern:

$$S = r\sigma_b$$

where σb is the standard deviation of background concentrations and *r* is the specified proportion.

This choice for a substantial difference is closely related to the use of a percentile of the background distribution discussed in Section A.3.2.

Areas with relatively high mean background concentrations generally also have high variance of background. Values of *r* less than 1 may require a high number of samples, for the reasons noted in Section A.2.2.

A.2.4 Proportion of Preliminary Remediation Goal

The concept of calculating risk-based soil concentrations to serve as reference points for establishing site-specific cleanup levels was introduced in RAGS. If a preliminary remediation goal (PRG) is available for the contaminant of concern, the value of S may be based on a proportion of the PRG:

$$S = r \cdot PRG$$

A proportion less than 1 may be required, because the total risk will be the sum of the incremental risk due to S plus the risk due to background concentrations of the contaminant. If the PRG is less than the mean or standard deviation of background, a high number of samples may be required for conclusive test results.

A.2.5 Proportion of Soil Screening Level

If a PRG is not available for the contaminant of concern, a risk-based value of S may be based on the soil screening level (SSL) for the contaminant.[4]

$$S = r \cdot SSL$$

SSLs are based on a 10^{-6} individual risk for carcinogens and a hazard quotient of 1 for noncarcinogens. SSLs were established to identify the lower bound of the range of risks of interest in decision making, and are not cleanup goals. SSL target risks should be adjusted to reflect established cleanup level targets. Again, a proportion less than 1 may be required, because the total individual risk will be the sum of the incremental risk due to S plus the risk due to background concentrations of the contaminant. If the (adjusted) SSL is less than the mean or standard deviation of background, a high number of samples may be required for the background comparison.

A.3 Statistical Tests and Confidence Intervals for Background Comparisons

This section provides supplementary material on the use of hypothesis tests and confidence intervals for conducting background comparisons. The science of statistics is often divided into two parts: estimation theory and hypothesis testing. Estimation theory includes the calculation of confidence intervals as estimates for population parameters, while hypothesis testing focuses on the use of statistical tests to accept or reject hypotheses concerning these parameters. Although only the use of hypothesis tests has been discussed in the main text, the one-to-one correspondence between hypothesis tests for Δ conducted at level α and the estimated $100(1-\alpha)$ percent confidence interval for Δ permits the use of either method to conduct a background comparison. While the emphasis of this section is technical in nature, mathematical proofs of results have been omitted.

When using Test Form 1, a one-sided, level-α hypothesis test of the null hypothesis $\Delta \leq 0$ will only reject the null hypothesis if we conclude that Δ is significantly greater than 0 by comparing the test statistic to the tabulated critical value. The critical value is selected to ensure that the probability that the test statistic will exceed the critical value by chance alone is less than α. A similar conclusion is reached when the lower bound of the one-sided, $100(1-\alpha)$ percent confidence interval for Δ is greater than zero. There are two ways to reach the same conclusion, that Δ is significantly greater than zero. A two-sided confidence interval for Δ is often more useful than one-sided confidence intervals to summarize the information about Δ that is contained in the data. In this case, a two-sided, $100(1-\alpha)$ percent confidence interval for Δ will correspond to a one-sided, level-$\alpha/2$ hypothesis test for Δ.

A.3.1 Comparisons Based on the t-Test

Background comparisons based on the t-test rely on the assumption of a normal distribution for the data, or for a transformation of the data. Hypotheses are tested using the t-statistic, which follows the Student t-distribution. Similar results are obtained by estimating a confidence interval for $\Delta = \mu_y - \mu_x$, where μ_y is the mean concentration in the contaminated area and μ_x is the mean background concentration.

NORMAL THEORY, CASE 1: EQUAL BUT UNKNOWN VARIANCES[5]

For simplicity, we first assume that the site data (Y_1, …, Y_n) and background data (X_1, …, X_m) are independent random samples from normal distributions with the same variance, σ^2, but with different means, μ_y and μ_x, respectively:

$$Y_j \sim N [\ \mu_y, \sigma^2\ ]$$

and

$$X_j \sim N [\ \mu_x, \sigma^2\ ].$$

In this case, the test statistic for the two-sample t-test is based on the difference in the estimated means, M_y and M_x, where

$$M_y = \Sigma\ Y_j /n \sim N [\ \mu_y, \sigma^2/n\ ]$$

and

$$M_x = \Sigma\ X_j / m \sim N [\ \mu_x, \sigma^2/m\ ].$$

A pooled estimate for σ^2, the common variance of the distributions, is

$$s_p^2 = [\ \Sigma\ (Y_j - M_y)^2 + \Sigma\ (X_j - M_x)^2\ ] / (n + m - 2).$$

The test statistic for conducting a t-test using Background Test Form 1 is

$$t_1 = (M_y - M_x) / s^*$$

where

$$s^* = s_p(1 / n + 1 / m\ )^{½}.$$

In Background Test Form 1, the test statistic t_1 has the standardized Student-t distribution with n+m-2 degrees of freedom if $\mu_y = \mu_x$ ($\Delta = 0$). Let $t_{1-\alpha}$ represent the $100(1-\alpha)^{th}$ quantile of the Student t-distribution with n+m-2 degrees of freedom. The value $t_{1-\alpha}$ is the critical value for the test. If the test statistic t_1 exceeds the critical value $t_{1-\alpha}$, the null hypothesis in Background Test Form 1 (H_o: $\Delta \leq 0$) may be rejected with 100(1-α) percent confidence.

The test statistic for conducting a two-sample t-test using Background Test Form 2 is

$$t_2 = (M_x + S - M_y)/s^*$$

where the quantity S is a substantial difference. The test statistic t_2 has a standard Student-t distribution with n+m-2 degrees of freedom when $\mu_S = \mu_B + S$. If the test statistic t_2 exceeds the critical value $t_{1-\alpha}$, then the null hypothesis in Background Test Form 2 (H_o: $\Delta > S$) may be rejected with 100(1-α) percent confidence.

A 100(1-α) percent confidence interval for Δ is an interval denoted as (Δ_1, Δ_2) that satisfies the requirement

$$Pr\{\Delta_1 \leq \Delta \leq \Delta_2\ \} \geq 1 - \alpha$$

Here Δ_1 represents the lower limit of the confidence interval, and Δ_2 represents the upper limit of the confidence interval. Although one-sided hypothesis tests were considered above, the desired confidence interval is two-sided and symmetric, meaning that there is a probability of α/2 that Δ will be below this interval and a probability of α/2 that it will be above this interval.

If the lower limit of a 100(1-α) percent confidence interval for Δ is greater than zero, then the mean in the contaminated area is significantly greater than the background mean. This means that a one-sided, level-α/2 test of the null hypothesis H_0: $\Delta \leq 0$ (Test Form 1) will reject the null hypothesis. Similarly, if the upper limit of a 100(1-α) percent confidence interval for Δ is less than *S*, then the difference between the mean in the contaminated area and the background mean is significantly less than a substantial difference. This means that a one-sided, level-α/2 test of the null hypothesis H_0: $\Delta > S$ (Test

Form 2) will reject the null hypothesis.

A symmetric confidence interval for the difference $\Delta = \mu_y - \mu_x$ is constructed using $t_{1-\alpha/2}$, which represents the $100(1-\alpha/2)^{th}$ quantile of the Student t-distribution with n+m-2 degrees of freedom. A $100(1-\alpha)$ percent confidence interval for Δ has the form (Δ_1, Δ_2), where the lower bound is

$$\Delta_1 = (M_y - M_x) - t_{1-\alpha/2}s^*$$

and the upper bound is

$$\Delta_2 = (M_y - M_x) + t_{1-\alpha/2}s^*.$$

Although the distribution of the test statistic for the two-sample Student t-test is derived based on the assumption of normal distributions with equal variances, the test is robust and has demonstrated good performance when the variances are unequal, and when the population distributions are not normal. However, the estimates M_y, M_x and s_p^2 are sensitive to outliers in either data set. If either or both data sets contain non-detects, then the test will be sensitive to most common methods of handling these values. Confidence intervals derived using the two-sample test statistic are expected to have similar properties.

NORMAL THEORY, CASE 2: UNEQUAL, UNKNOWN VARIANCES[6]

Now assume that the site data $(Y_1, \ldots, Y_n)$ and background data $(X_1, \ldots, X_m)$ are independent random samples from normal distributions with different means, μ_y and μ_x, and different variances, σ_y^2 and σ_x^2, respectively:

$$Y_j \sim N[\ \mu_y, \sigma_y^2\ ]$$

and

$$X_j \sim N[\ \mu_x, \sigma_x^2\ ].$$

Estimates for the sample variances are

$$s_y^2 = \Sigma(Y_j - M_y)^2 / (n - 1)$$

and

$$s_x^2 = \Sigma(X_j - M_x)^2 / (m - 1).$$

An estimate of the approximate degrees of freedom is

$$\nu = \tau^2/b$$

where

$$\tau = s_y^2/n + s_x^2/m$$

and

$$b = (s_y^2/n)^2 / (n - 1) + (s_x^2/m)^2 / (m - 1).$$

A symmetric confidence interval for the difference $\Delta = \mu_y - \mu_x$ is constructed using the Student t-distribution with ν^* degrees of freedom, where ν^* is the closest positive integer to ν. Let $t_{1-\alpha/2}$ represent the $100(1-\alpha/2)^{th}$ quantile of this t-distribution with ν^* degrees of freedom. An approximate $100(1-\alpha)$ percent confidence interval for Δ has the form (Δ_1, Δ_2), where the lower bound is

$$\Delta_1 = (M_y - M_x) - t_{1-\alpha/2}\tau^{1/2}$$

and the upper bound is

$$\Delta_2 = (M_y - M_x) + t_{1-\alpha/2}\tau^{1/2}$$

A.3.2 Comparisons Based on the Wilcoxon Rank Sum Test

The Wilcoxon Rank Sum (WRS)[7] test is a nonparametric test for testing whether there is a difference between the site and background population distributions. The WRS test examines whether measurements from one population tend to be consistently larger (or smaller) than those from the other population. The test determines which is the higher distribution by comparing the relative ranks of the two data sets when the data from both sources are sorted into a single list. One assumes that any difference between the site and background concentration distributions represents a shift of the site concentrations to higher values due to the presence of contamination in addition to background. The WRS test is most effective when contamination is spread throughout a site.

Two assumptions underlying the WRS test are:

1) Samples from the background and site are independent, identically distributed random

samples; and

2) Each measurement is independent of every other measurement, regardless of the set of samples from which it came.

The WRS test assumes that the distributions of the two populations are identical in shape (variance), although the distributions need not be symmetric.

The WRS test has three advantages over the t-test for background comparisons:

1) The two data sets are not required to be from a known type of distribution. The WRS test does not assume that the data are normally or log-normally distributed, although a normal distribution approximation often is used to determine the critical value for the test for large sample sizes.

2) The WRS test is robust with respect to outliers because the analysis is conducted in terms of ranks of the data. This limits the influence of outliers because a given data point can be no more extreme than the first or last rank.

3) The WRS test allows for non-detect measurements to be present in both data sets. The WRS test can handle a moderate number of non-detect values in either or both data sets by treating them as ties.[8]

Theoretically, the WRS test can be used with up to 40 percent or more non-detect measurements in either the background or the site data. Such a high proportion of non-detects indicates that there will be a large number of ties. In this case, the simple expediency of assigning all ties the same ranks may not be adequate. More specific procedures have been developed to address data sets with a large number of ties.[9] If more than 40 percent of the data from either the background or site are non-detect values, the WRS test should not be used.

The WRS test may be applied to both forms of background test, *indistinguishable from background* or *exceed by more than a substantial difference*. In either form of background test, the null hypothesis is assumed to be true unless the evidence in the data indicates that it should be rejected in favor of the alternative.

The WRS test for Background Test Form 1 is applied as outlined in the following steps. The site and background measurements are ranked in a single list in increasing order from 1 to N, where N = m + n. All tied values are assigned the average of the ranks for that group of measurements. All non-detect values are considered as ties and are assigned an average rank (if there are a total of *t* non-detects, they all are assigned rank (t+1)/2, which is the average of the first *t* integers).

The sum of the ranks of the site measurements (W_y) and the sum of the ranks of the background measurements (W_x) are sufficient statistics for the test, where $W_y + W_x = N(N + 1)/2$. The sum of the ranks of the site measurements (W_y) is the test statistic used for Background Test Form 1. To conduct the test, W_y is compared with w_α, which is the critical value for a level-α WRS test for the appropriate values of n and m[10]. If W_y exceeds the critical value for the test, the null hypothesis that the site is indistinguishable from background ($\Delta \leq 0$) may be rejected with 100(1-α) percent confidence.

The WRS test for Background Test Form 2 is applied as outlined in the following steps. First, the background measurements are adjusted by adding the substantial difference S to each measured value[11]. Second, the S-adjusted background data and the site data are ranked in a single list in increasing order from 1 to N. Finally, all tied values are assigned the average of the ranks for that group of measurements.

The sum of the ranks of the S-adjusted background measurements (W_{x+S}) is the test statistic for Background Test Form 2. If W_{x+S} is greater than the critical value for the test, w_α, the null hypothesis that the site exceeds background by more than a substantial difference ($\Delta > S$) is rejected at the 100(1-α) percent confidence level.

Nonparametric confidence intervals for Δ are derived based on the Mann-Whitney form of the WRS test (Section 5.3.2). The Mann-Whitney test statistics are computed from the set of all possible differences between the site and background data sets:

$$\{Y_i - X_j, I = 1, \ldots, n; j = 1, \ldots, m\}.$$

There are *n* times *m* possible differences in this set, so a computer program may be required to perform the necessary calculations. Let the symbol Z_r (r = 1, … , nm) represent the *r*th-ranked difference in the ordered set of all possible differences between the site and background data. A symmetric nonparametric confidence interval for Δ is constructed using the *k*th-smallest ranked difference (Z_k) and the *k*th-largest ranked difference (Z_{nm-k+1}) in the set of all possible differences, where k depends on n, m and α[12]. Thus, a $100 \times (1-\alpha)$ percent confidence interval for Δ is a closed interval of the form

$$(\Delta_1, \Delta_2) = (Z_k, Z_{nm-k+1})$$

with

$$k = w_{\alpha/2} - n(n + 1)/2.$$

Here, as noted above for the WRS test, $w_{\alpha/2}$ is the tabulated critical value for a level-$\alpha/2$ WRS test for the appropriate values of n and m. This confidence interval satisfies the requirement

$$\Pr\{ \Delta_1 \leq \Delta \leq \Delta_2 \} \geq 1 - \alpha.$$

APPENDIX NOTES

1. U.S. Environmental Protection Agency (EPA). July 2000. *Guidance for Data Quality Assessment: Practical Methods for Data Analysis, EPA QA/G-9, QA00 Version*. Quality Assurance Management Staff, Washington, DC. EPA 600-R-96-084. Available at http://www.epa.gov/quality/qa_docs.html.

2. U.S. Environmental Protection Agency (EPA). 1994. *Guidance for the Data Quality Objectives Process, EPA QA/G-4*, EPA 600-R-96-065. Washington DC.

3. U.S. Environmental Protection Agency (EPA). 1989. *Statistical Methods for Evaluating the Attainment of Cleanup Standards* Volume 3, subtitled *Reference-Based Standards for Soils and Solid Media*, EPA 230-02-89-042. Washington DC.

4. U.S. Environmental Protection Agency (EPA). 1996. *Soil Screening Guidance: Technical Background Document*, EPA 540-R-95-128.

5. Zwillinger, D. and S. Kokoska. 2000. *CRC Standard Probability and Statistics Tables and Formulae*, Chapman and Hall/CRC Press, New York, Section 9.6.2.

6. Zwillinger and Kokoska, *Op. Cit.*, Section 9.6.3.

7. The WRS test is also called the Mann-Whitney test, which is mathematically equivalent to the WRS test. Sometimes, the combined name is used: Wilcoxon-Mann-Whitney test. See Section 5.3.2.

8. The Gehan form of the WRS test should be considered if there are many non-detect values with different detection levels.

9. If there are many ties, see instructions for applying the WRS test in Conover, W. J., *Practical Nonparametric Statistics, 2nd Ed.*, John Wiley & Sons, Inc., New York, NY, 1980.

10. Critical values for the WRS test are available in many published texts and reference books. Two sources are Conover, W.J., *Practical Nonparametric Statistics, 2nd Ed.*, John Wiley & Sons, NY, 1980; and *CRC Standard Probability and Statistics Tables and Formulae*, D. Zwillinger and S. Kokoska, Chapman and Hall/CRC Press, Boca Raton, Florida, 2000.

11. Conover, *Op. Cit.*, p. 223, Equation 8.

12. Conover, *Op. Cit*., p. 224.