HANDBOOK FOR STATISTICAL ANALYSIS OF ENVIRONMENTAL BACKGROUND DATA

Prepared by:

SWDIV and EFA WEST
of
Naval Facilities Engineering Command

July 1999

CONTENTS

FIGURES

TABLES

BOXES

ACKNOWLEDGMENTS

This handbook was prepared with technical assistance from the following individuals and firms. Nancy C. Hassig, Ph.D., Statistics Resources, Battelle Pacific Northwest Division, Richland, WA was the project manager. She also supervised and assisted with the development/production of graphics and wrote portions of the text. The principal author of the handbook was Richard O. Gilbert, Ph.D., Statistics Resources, Battelle Washington Office, Washington, D.C. Derrick (Rick) J. Bates, M.S., Statistics Resources, Battelle Pacific Northwest Division, produced the graphs and figures. Mary H. (Nell) Cliff, also with Statistics Resources, Battelle Pacific Northwest Division, incorporated the graphics and text into the final document. Tom Grieb, M.S., Tetra Tech Environmental Management Inc. provided an initial draft document that aided in the development of this handbook.

Most of the statistical methods in the handbook were drawn from the U.S. Environmental Protection Agency (EPA) document, "Guidance for Data Quality Assessment, Practical Methods for Data Analysis," EPA QA/G-9, EPA/600/R-96/084, developed under the direction of Dr. John Warren, Ph.D., of the U.S. EPA, Quality Assurance Division, Washington, D.C.

This document was prepared under the supervision and contribution of the Naval Facilities Engineering Command, Southwest Division, representatives Dennis Askvig of Engineering Field Division, Southwest, and Camille Garibaldi (now with the Federal Aviation Administration), Gilbert Rivera and Kenneth Spielman of Engineering Field Activity, West. Their guidance, support, and review were important contributions to the development of this handbook. Special acknowledgments are also due to those individuals who reviewed the final draft version of this handbook.

Questions concerning this document should be directed to:

Dennis Askvig, Code 03
Southwest Division, Naval Facilities Engineering Command
1220 Pacific Highway
San Diego, CA 92132
(619) 532-2510
email: askvigdw@efdsw.navfac.navy.mil

Gilbert Rivera, Code 7023
Engineering Field Activity, West
Naval Facilities Engineering Command
900 Commodore Drive
San Bruno, CA 94066
email: riveraga@efawest.navfac.navy.mil

Ken Spielman, Code 70233
Engineering Field Activity, West
Naval Facilities Engineering Command
900 Commodore Drive
San Bruno, CA 94066
(650) 244-2539
email: spielmankh@efawest.navfac.navy.mil

ACRONYMS AND ABBREVIATIONS

BRAC	Base Realignment and Closure
CF	Cumulative Frequency
COPC	Contaminants of Potential Concern
CV	Coefficient of Variation
$DCGL_{EMC}$	Derived Concentration Guideline Limit for the EMC
DL	Detection Limit
DQO	Data Quality Objectives
DQA	Data Quality Assessment
EMC	Elevated Measurement Comparison
EPA	U.S. Environmental Protection Agency
IRP	Installation Restoration Program
MARSSIM	Multi-Agency Radiation Survey and Site Investigation Manual
ND	Non-detect
QA	Quality Assurance
QAPP	Quality Assurance Project Plan
SD	Standard Deviation
SQL	Sample Quantitation Limit
SRS	Simple Random Sampling
WRS	Wilcoxon Rank Sum test

EXECUTIVE SUMMARY

This document provides step-by-step instructions for conducting graphical and statistical data analyses and tests of hypotheses to identify contaminants of potential concern (COPC) at Navy installations throughout California. The methods described in this handbook are provided to implement the guidance in the Navy document, "Procedural Guidance for Statistically Analyzing Environmental Background Data" (Navy 1998). The Navy intends to implement the guidance in these two documents at all California installations. Such implementation will promote consistency throughout the Navy's Installation Restoration Program (IRP) and Base Realignment and Closure (BRAC) program to increase public and regulatory confidence in Navy cleanup activities.

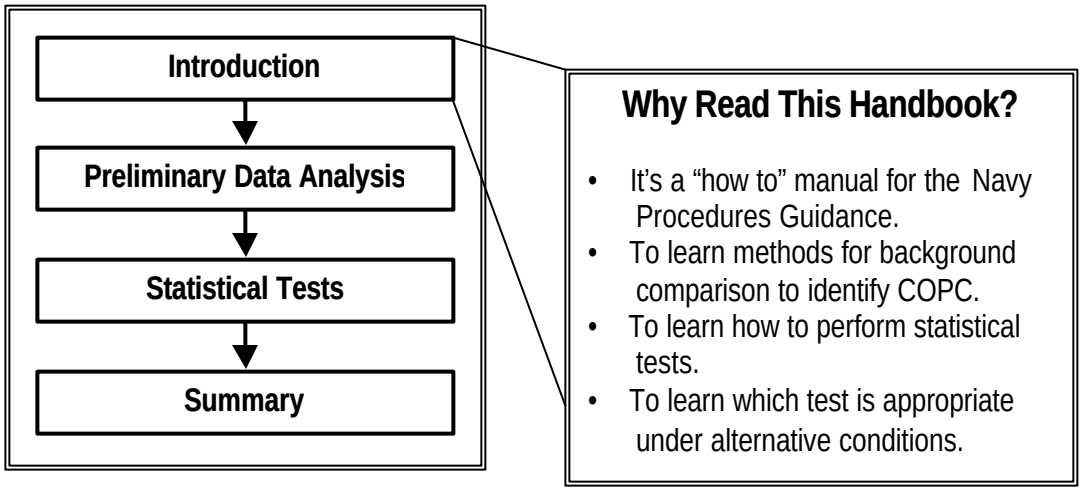
This document should be used *hand-in-hand* with the procedural guidance document (Navy 1998) that

- briefly reviews statutory requirements, regulations, and risk assessment guidance
- suggests approaches for developing representative background data sets
- provides an overview of the Data Quality Objectives (DQO) planning process and the Data Quality Assessment (DQA) process developed by the U.S. Environmental Protection Agency (EPA)
- provides a flowchart (Figure 11) to help project teams decide what statistical methods and tests should be used to identify the COPC.

The methods and statistical tests recommended in the flowchart (Figure 11 in Navy 1998) are discussed in detail in this handbook. All required statistical tables for performing the tests are provided, as are references to pertinent statistical literature for additional information. In addition to describing statistical tests for COPC, this handbook also describes graphical plots, descriptive statistics, tests for data outliers, and tests to evaluate the form of the distribution of data sets. These analyses are used to describe and communicate the information in site and background data sets, to look for data that may be errors and hence should be discarded, and to help decide which statistical tests for COPC are preferred. Formulas or tables are also provided to determine the number of samples that should be taken to conduct the statistical tests for COPC.

The data analysis and statistical testing methods described in this handbook closely follow EPA guidance in "Guidance for Data Quality Assessment, Practical Methods for Data Analysis," EPA (1996) developed by the EPA Quality Assurance Division.

1.0 INTRODUCTION



The purpose of this handbook is to provide Navy environmental restoration project teams with detailed instructions for selecting and conducting graphical and statistical analyses of environmental contaminant data. Such data, in turn, will help determine if a chemical is a contaminant of potential concern (COPC) to the health of humans or to the environment. This handbook implements and illustrates the statistical methods that are recommended in the Navy document "Procedural Guidance for Statistically Analyzing Environmental Background Data" (Navy 1998). The handbook will be implemented at all California Navy installations to promote consistency throughout the Navy Installation Restoration Program (IRP) and its Base Realignment and Closure (BRAC) activities.

This handbook was written assuming that the 7-step Data Quality Objectives (DQO) planning process (EPA 1993, 1994) will be used to determine the type, quantity and quality of environmental data needed to support COPC decisions at Navy sites. Proper use of the DQO process will provide the scientific foundation for defensible decision-making by helping to assure that representative field samples are collected at appropriate locations and times, that appropriate graphical and statistical analyses of the resulting data are conducted, and that appropriate interpretations of the data and statistical procedures are made. Additional information on the DQO process as it should be applied to determining the COPC at Navy sites is provided in Navy (1998).

The target audience for this handbook includes scientists who conduct risk assessments and background studies, scientists in regulatory agencies who review these risk assessments and studies, and Navy and regulatory remedial project managers and engineers who make decisions regarding the Navy environmental programs. The statistical methods described here are consistent with those described in U.S. Environmental Protection Agency (EPA) publications and guidance documents. [EPA (1994b, 1996, 1997, 1998), MARSSIM (1997)].

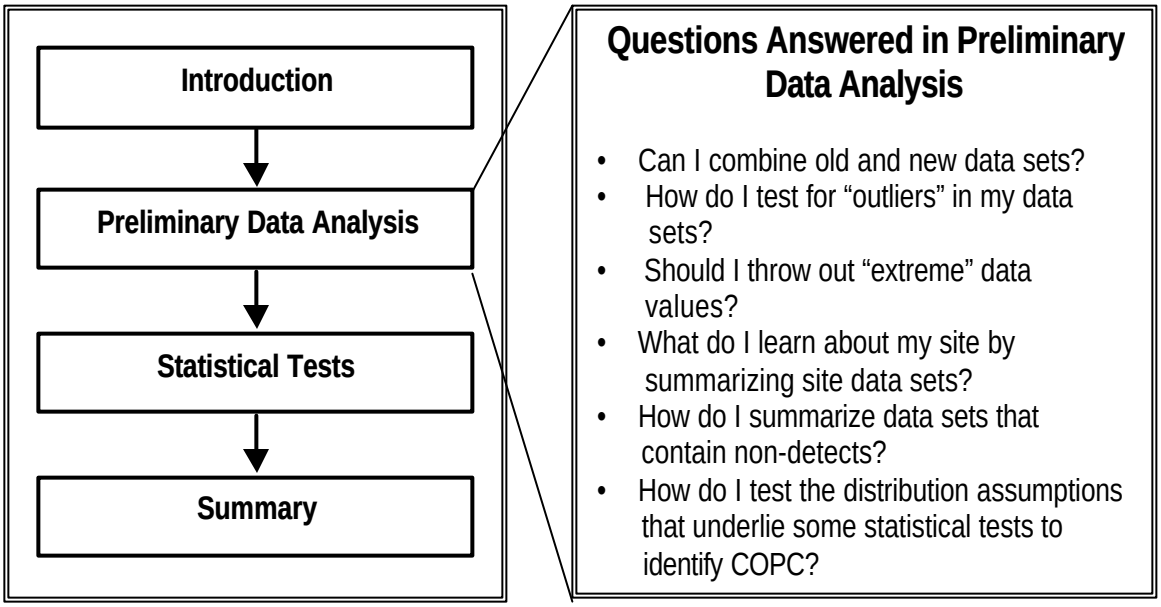
The Navy (1998) document emphasizes the importance of developing a valid set of background data before conducting statistical tests to determine COPC. Background data are those data that are collected from areas that have not been impacted by past or current activities or operations at the Navy site of interest. When comparing site data to background data, it is necessary to recognize the two categories of background substances: *naturally occurring* (those substances not contributed by human activities) and *anthropogenic* (natural and man-made substances that arise from human activities not related to site activities). As stated in Navy (1998): "It is essential that naturally occurring and anthropogenic background levels be established to accurately identify chemicals of concern and to estimate site risks specifically associated with Navy releases."

In this handbook, it is assumed that an appropriate background area (or areas) has been identified for comparison with the Navy site and that background concentrations (from both natural and anthropogenic substances) at the site are at the same level as in the background area. If anthropogenic background is at higher levels on the Navy site than in the background area, the magnitude of this difference must be determined so the higher anthropogenic levels at the site are not mistaken for site releases. It is the amount of increase in chemical levels that arises from *site activities* that is of interest.

This handbook describes and illustrates how to conduct several statistical tests for COPC. However, before a test procedure can be selected and used, it is necessary to evaluate the quality and quantity of any suitable data that are currently in hand. This task is accomplished using the preliminary data analyses described in Section 2.0 that are key elements of the Data Quality Assessment (DQA) process (EPA 1996, 1998). The DQA process evaluations provide information for deciding if the data are of the required type, quantity and quality (as specified during the DQO process used to plan the study) and if the assumptions that underlie the selected statistical test (or tests) to determine if a chemical is a COPC are valid. A full discussion and illustration of the DQA process and the statistical analysis and test procedures needed to implement the process are described in EPA (1996, 1998). The *DataQUEST* software (EPA 1997) can be used to perform most of the analyses described in these two EPA documents and in this handbook. *DataQUEST* may be downloaded from the EPA Quality Assurance Division Internet home page: http://es.epa.gov/ncerqa/qa/index.html.

This handbook is organized as follows. Chapter 2 discusses preliminary data analyses that consist of computing summary (descriptive) statistics and plotting graphical visual aids for evaluating the quality and quantity of the site and background data. These analyses are conducted to evaluate the quality of the data for determining COPC and to help the user identify and understand any problems with the data that may affect how the data are statistically analyzed. Chapter 3 provides case studies and examples to describe and illustrate statistical hypothesis tests that can be used to evaluate whether concentrations of contaminants in soil at Navy facilities exceed those in a suitable background area. The assumptions, advantages, and disadvantages of the various tests are provided as an aid in selecting the most appropriate test. Chapter 4 provides summary comments and discussions on the use of this handbook. A glossary of key words and phrases is also provided. Chapter 7 is a set of tables required by the statistical tests.

2.0 PRELIMINARY DATA ANALYSES



2.1 Introduction

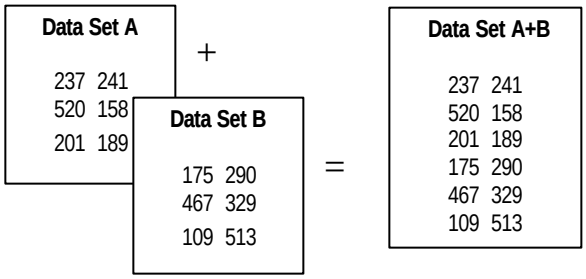
When the DQO planning process is complete, appropriate background and Navy site field samples are collected at locations and times according to the specified sampling design as determined via the DQO process. After the collected samples have been processed and measured for the specified constituents, an evaluation of these measurements must be made to assure that they are the type, quantity, and quality that was specified during the DQO process and that are required by the statistical tests selected for determining the COPC. This evaluation should be conducted using the DQA process, that consists of the following steps:

1. Review of DQO (output of each step of the DQO process) and sampling design
2. Conduct preliminary data review
3. Select the statistical test
4. Verify the assumptions
5. Draw conclusions from the data.

In this chapter, we focus on Step 2 and Step 4 of the DQA process. Step 2 consists of (1) reviewing quality assurance (QA) reports that describe the data collection, measurement, and reporting process that was used for the site and background data, and (2) computing descriptive statistics and graphical pictures of the data to quantify and visualize the mean and variability of the background and site measurements. Step 4 consists of verifying the validity of the assumptions that underlie the statistical tests selected for identifying the COPC. These tests should have been tentatively selected during the DQO process. Some statistical tests require the measurements to have a normal (Guassian) distribution; all tests require that any measurements that are errors be identified and removed.

The review of QA reports is beyond the scope of this handbook, but is briefly discussed in EPA (1996, p. 2.1-1). Before conducting descriptive statistics and plotting graphical analyses it is important to verify that all appropriate historical data have been located and combined with the newly collected data and that data sets contain no measurements that are mistakes or errors. Section 2.2 discusses how to determine if data sets should be combined. Section 2.3 shows how to conduct statistical analyses to look for data outliers, those measurements that are so large as to suggest they may be mistakes and should be discarded. Sections 2.4 and 2.5 describe recommended descriptive summary statistics and graphical data analysis procedures, respectively, including cases where the data set contains non-detects. Section 2.6 shows how to test whether the site or background data sets are normally distributed, an assumption that underlies some statistical tests in Chapter 3 to determine COPC.

2.2 Combining Data Sets



Should Data Sets be Combined?

Combining two or more data sets to form a larger data set may improve the ability of statistical tests to detect when a contaminant is a COPC. For example, soil samples may have been collected and measured for the same suite of chemicals at several different times in the land area of concern at a Navy site. Pooling the data will increase the number of samples available for conducting a statistical test for a COPC and could increase the chances the test result will be accurate. However, an inappropriate combining of data sets can have the opposite effect. This section provides guidance on some questions that should be considered before pooling data sets.

Before data sets are combined, it is necessary to carefully define the target population of current interest for determining if the chemical of interest is a COPC. The target population is that set of environmental space/time units within spatial and time boundaries for which we wish to decide whether the chemical is a COPC. Each data set being pooled together must consist of representative data from the target population of *current* interest. If one data set was obtained from "Site A" at a Navy site before fill dirt was placed on the site, whereas a second data set was obtained from "Site A" after fill dirt was added, the concentrations of the chemical of interest may have changed quite drastically. That is, the underlying population of concentrations of the chemical of interest may now be quite different. Furthermore, neither target population may be the one that is currently present at Site A because recent site operations may have added the chemical of interest to the surface soil.

Ideally, the data sets being considered for pooling should have been obtained using the same sampling design that was applied to the same area of land. For example, it may not be a good idea to pool data collected along a straight line in one corner of the site with data collected using simple random sampling over the entire site. Concentration levels of the chemical could have been much higher in the area where the samples were collected along a straight line. Similarly, data collected at locations determined by expert judgment and pooled with data collected on a grid could lead to unrepresentative data for the site as a whole. Of course, if good evidence indicates the concentrations of the chemical are about the same over the entire site, the sample collection locations will not be a critical issue of concern. However, that assumption should not be made without substantial evidence that it is true.

It is also important to verify that measurements in all the data sets being considered for pooling have acceptable quality for the purpose at hand. For example, the detection limits, quantitation limits, and measurement biases should all be sufficiently low, and an adequate

number of blank and replicate samples should be taken to check for the magnitude of bias and variability. Furthermore, the same sample collecting, compositing, handling, and measuring methods should have been used for all the data sets that are being considered for pooling. If not, the burden of proof must show that any such differences in the data sets will not have an effect on the decisions made on the basis of the pooled data.

Graphical displays and statistical analysis methods should also be used to assure whether the data sets have clearly different amounts of scatter (variance) or different average concentrations. If so, pooling the data may not be warranted. Graphical methods, such as histograms, box plots, and probability plots (described and illustrated in Section 2.5) may be applied to each individual data set to look for differences. If only two data sets are being considered for pooling, the Wilcoxon Rank Sum test or the Gehan test (Gehan 1965), described in Sections 3.6 and 3.7, respectively, may be used to look for differences in the *medians* (defined in Box 2.1) of the two data sets, if the variances of the sets are approximately equal. Differences in the *means* of data sets that have a bell-shaped normal distribution may be tested using the two-sample t test or the Satterthwaite two-sample t test described in Sections 3.8 and 3.9, respectively. The Satterthwaite test is used if the variances are believed to differ. Furthermore, differences in the variance (defined in Box 2.1) of measurements for the two data sets that have a normal distribution (with possibly different means) could be tested using the F test described in EPA (1996, Box 4.5-2) and Iman and Conover (1983, page 275). Alternatively, the Squared Ranks Test of variances described in Conover (1980, page 239) may be used to test for equality of variances. This test may be used regardless of the shape of the data distributions.

If more than two data sets are being considered for pooling, the Kruskal-Wallis test (Gilbert, 1987, page 250; Conover 1980, page 229) may be used to look for differences in medians. A test for equal variances of more than two data sets is provided in Conover (1980, page 241). Both of these tests may be applied regardless of the shape of the underlying distribution.

2.3 Descriptive Summary Statistics

	n	ND	Mean	SD	Max
Al	202	0	6639	5625	39000
As	319	242	3	5	62
Be	202	63	0.3	0.4	5.9
Cr	332	6	12	13	116
Ni	250	119	5	19	224

What is the mean (central tendency)?
What is the standard deviation (spread of data)?
What are the maximum and minimum data values?

This section defines and describes how to compute descriptive summary statistics for the Navy site and background data sets as part of a preliminary data review. These descriptions, in conjunction with graphical plots discussed in Section 2.5, should be conducted to develop an understanding of the range, variability, and shape of the underlying probability distribution of the measurements, as well as the number of non-detects and possible outliers that are present. This information is needed to help determine the quality of the data sets and how the data should be statistically analyzed. This preliminary data review is needed to decide which statistical test(s) for COPC should be conducted.

An assumption that underlies conducting statistical tests is that measurements made from samples collected at a study site, be it the Navy site or the background area, are representative of the underlying population of all possible measurements for the chemical of interest at the study site. This assumption means the locations selected for collecting soil samples must yield representative measurements of the field population. Moreover, the methods used to collect, transport, prepare, and measure the soil samples must not introduce any bias into the measurements. If representative measurements are not obtained, the statistical tests used to decide which chemicals are COPC can be very misleading.

The best way to assure that representative sampling locations are selected is to determine the locations using a probability-based sampling design strategy. Two such designs are simple random sampling and systematic sampling. If systematic sampling is used, sample locations could be at the nodes of a square or triangular grid system that is placed at a random starting place in the area to be sampled. These and other designs are discussed in EPA (1999) and Gilbert (1987).

An additional assumption is that data sets do not contain spurious measurements. Such measurements can occur because of mistakes and errors during the sample collecting, handling, and measuring processes. Statistical tests for detecting outliers are provided in Section 2.4.

In Section 2.3.1, we consider descriptive summary statistics for cases where data sets do not contain any non-detects, that is, measurements that are below some quantitative upper limit, such as the detection limit or the quantitation limit. In Section 2.3.2, we consider the case of data sets that contain non-detects.

2.3.1 Data Sets with No Non-Detects

Data Sets with No Non-Detects

Concentrations of Copper in Soils (mg/kg)	Descriptive Statistics
7.7	
10.7	Mean = 28.6
14.3	Median = 18.3
18.3	Std. Dev. = 22.6
35.5	Min = 7.7
44.1	Max = 69.8
69.8	

Box 2.1 lists and defines the descriptive summary statistics that should be computed for the Navy site and background data sets. The number of measurements in a data set is denoted by n. The n measurements are denoted by x_1, x_2 ,…, x_n. Examples that show how to calculate the descriptive summary statistics are provided in Box 2.2.

Box 2.1. Descriptive Summary Statistics for Data Sets with No Non-Detects

Descriptive Statistics	Definitions and Computation
Arithmetic Mean ($\bar{x}$)	$\bar{x} = (x_1 + x_2 + \ldots + x_n) / n$
Median (when n is an odd integer)	The middle value of the n measurements after they are arranged in order of magnitude from smallest to largest
Median (when n is an even integer)	The arithmetic average of the middle two of the ordered measurements
p^{th} sample percentile	The value (not necessarily an observed measurement) that is greater than or equal to p% of the values in the data set and less than or equal to (1-p)% of the data values, where $0 < p < 1$. Compute $k = p(n + 1)$, where n is the number of measurements. If k is an integer, the p^{th} percentile is the k^{th} largest measurement in the ordered data set. If k is not an integer, the p^{th} percentile is obtained by linear interpolation between the two measurements in the ordered data set that are closest to k.
Range	The maximum measurement minus the minimum measurement
Interquartile range	The 75^{th} sample percentile minus the 25^{th} sample percentile
Sample Standard Deviation (SD)	A measure of dispersion (spread or variation) of the n measurements in a data set that is computed as follows: $SD = \{[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2] / (n - 1)\}^{1/2}$
Sample Variance	The sample variance is the square of the sample SD, that is, Sample Variance = $(SD)^2$.
Coefficient of Variation (CV)	The CV is a measure of relative standard deviation that is computed as follows: $CV = s / \bar{x}$

Box 2.2. Examples of Descriptive Summary Statistics for Data Sets with No Non-Detects

Descriptive Statistics	Example Calculations
Arithmetic Mean ($\overline{x}$)	Suppose there are 5 data, say 50, 34, 52, 62, 60. Then the arithmetic mean is $\overline{x} = (50 + 34 + 52 + 62 + 60) / 5 = 51.6$
Median (when n is an odd integer)	For the 5 data (after being ordered from smallest to largest) 34, 50, 52, 60, 62, the median is 52.
Median (when n is an even integer)	Suppose there are 6 data, which when ordered from smallest to largest are 0.1, 0.89, 2.0, 3.01, 3.02, 4.0. Then the median is (2.0 + 3.01) / 2 = 2.50
p^{th} sample percentile	Suppose the data set (after being ordered) is 34, 50, 52, 60, 62, and we want to estimate the 60^{th} percentile, that is, p = 0.6. Now, k = 0.6 (5 + 1) = 3.6. Since k is not an integer, we linearly interpolate between the 3^{rd} and 4^{th} largest measurements, that is, the 0.60 sample percentile is 52 + 0.6 (60 – 52) = 56.8.
Range	For the data set 50, 34, 52, 62, 60 the range is 62 – 34 = 28.
Interquartile Range	The 75^{th} sample percentile of the (ordered) data set 34, 50, 52, 60, 62 is 60 + 0.5(62 – 60) = 61. The 25^{th} sample percentile is 34 + 0.5(50 – 34) = 42. Therefore, the interquartile range is 61 – 42 = 19
Sample Standard Deviation (SD)	The sample SD of the data set 50, 34, 52, 62, 60 is $SD = \{ [(50 – 51.6)^2 + (34 – 51.6)^2 + (52 – 51.6)^2 + (62 – 51.6)^2 + (60 – 51.6)^2] / 4\}^{1/2} = 11.08$
Sample Variance	The sample variance of the data set 50, 34, 52, 62, 60 is the square of the sample SD, that is, Variance = $(11.08)^2 = 122.77$.
Coefficient of Variation (CV)	The CV for the data set 50, 34, 52, 62, 60 is CV = 11.08 / 51.6 = 0.21 .

2.3.2 Data Sets That Contain Non-Detects

Data Sets with Non-detects

Concentrations of Copper in Soils (mg/kg)	Adjusted Descriptive Statistics
< 12.0	
< 12.0	Median = 18.3
14.3	Trimmed Mean = 22.7
18.3	Winsorized Mean = 24.0
35.5	Winsorized Std. Dev. = 32.7
44.1	
69.8	

Non-detects are measurements that the analytical laboratory reports are below some quantitative upper limits such as the detection limit or the limit of quantitation. Data sets that contain non-detects are said to be censored data sets.

The methods used to compute descriptive statistics when non-detects are present should be selected based on the number of non-detects and the total number of measurements, n (detects plus non-detects). If n is large (say, n > 25) and less than 15% of the data set are non-detects, the general guidance in Navy (1998) and EPA (1996) is to replace the non-detects with DL (Detection Limit), DL/2, or a very small value. The descriptive summary statistics in Box 2.1 may then be computed using the (now) full data set, although some of the resulting statistics will be biased to some degree. (The median, pth sample percentile, and the interquartile range may not be biased if the number of non-detects is sufficiently small.) The biases may be large, even though less than 15% of the measurements are non-detects, particularly if n is small, say n < 25.

If 15% to 50% of the data set are non-detects, the guidance offered in EPA (1996, 1998) and Navy (1998) is to forgo replacing non-detects with some value like the DL divided by 2, the DL itself, or a small value. Instead, one should consider computing the mean and standard deviation using the Cohen method or computing a trimmed mean or a Winsorized mean and standard deviation. These methods, as well as the Winsorized standard deviation, are defined and their assumptions, advantages, and disadvantages are listed in Box 2.3. Examples of computing the median, trimmed mean, the Winsorized mean and standard deviation are illustrated in Box 2.4. The Cohen method for computing the mean and standard deviation of a normally distributed set of data that contains non-detects is explained and illustrated in Box 2.5.

If more than 50% of the measurements in the data set are non-detects, the loss of information is too great for descriptive statistics to provide much insight into the location and shape of the underlying distribution of measurements. The only descriptive statistics that might be possible to compute are p^{th} percentiles for values of p that are greater than the proportion of non-detects present in the sample and when no non-detects are greater than the $k(n+1)^{th}$ largest datum, where k is defined in Box 2.1.

It must be noted that EPA (1996) cautions that no general procedures exist for the statistical analyses of censored data sets that can be used in all applications of statistical

analysis, that is, for all purposes, and that EPA guidelines should be implemented cautiously. EPA (1996) also suggests the data analyst should consult a statistician for the most appropriate way to statistically evaluate or analyze a data set that contains non-detects.

Akritas, Ruscitti, and Patil (1994) provide a review of the statistical literature that deals with the statistical analysis of censored environmental data sets. A review for those persons who are not so familiar with statistical methods is provided by Helsel and Hirsch (1992).

Box 2.3. Descriptive Statistics when 15% to 50% of the Data Set are Non-Detects (Gilbert 1987; EPA 1996)

Method	Assumptions	Advantages/Disadvantages
Median (when n is an odd or an even integer): Determine the median in the usual way as illustrated in Box 2.1	• The largest non-detect is less than the median of the entire data set (detects + non-detects), that is, there are no non-detects in the upper 50% of the measurements	Advantage: • A simple procedure Disadvantage: • The median cannot be determined, if the assumption is not true.
100p% Trimmed Mean: Determine the percentage (100p%) of measurements below the DL. Discard the largest np measurements and the smallest np measurements. Compute the arithmetic mean on the n(1-2p) remaining measurements.	• All non-detects have the same DL. • All detects are larger than the DL • The number of non-detects is no more than np. • The underlying distribution of measurements is symmetric (not skewed). • $0 < p < 0.50$.	Advantage: • Trimmed mean is not affected by outliers that have been trimmed from the data set. Disadvantages: • Cannot be used if the assumptions are not true.
Winsorized Mean ($\overline{x}_w$): If n′ non-detects are in the lower tail of a data set with n measurements (including non-detects). • Replace the n′ non-detects by the next *largest* detected datum. • Also replace the n′ largest measurements by the next *smallest* measurement. • Obtain the Winsorized Mean, $\overline{x}_w$, by computing the arithmetic mean of the resulting set of n measurements	• All non-detects have the same detection limit (DL). • All detects are larger than the DL • The underlying distribution of the measurements is symmetric (not skewed).	Advantage: • Winsorized mean is not affected by outliers that are among the largest measurements. Disadvantage: • Cannot be used if the assumptions are not true.

Winsorized Standard Deviation (s_w) Suppose n′ non-detects are in the lower tail of a data set with n measurements (detects plus non-detects). • Replace the n′ non-detects by the next *largest* detected datum. • Also replace the n′ largest measurements by the next *smallest* measurement. • Compute the standard deviation, s, of the new set of n measurements. • Compute $s_w = [s(n-1)]/(v-1)$, where $v = n - 2n'$ is the number of measurements not replaced during the Winsorization process.	• All non-detects have the same detection limit (DL). • All detects are greater than the DL. • The underlying distribution of the measurements is symmetric (not skewed). • The quantity v must be greater than 1.	Advantage: • If the measurements are normally distributed, then confidence intervals for the mean can be computed using the method in Gilbert (1987, page 180) Disadvantage: • Cannot be used if the assumptions are not true.
Cohen Method for the Mean and Standard Deviation. (See Box 2.5)	• All non-detects have the same DL • The underlying distribution of the measurements is normal. • Measurements obtained are representative of the underlying normal distribution.	Advantage: • Has good performance if the underlying assumptions are valid and if the number of samples is sufficiently large. Disadvantage: • The assumptions must be valid.
p^{th} Sample Percentile The pth sample percentile is computed as described in Box 2.1.	• All non-detects have the same DL. • All detects are greater than the DL. • The computed value of k (see Box 2.1) must be larger than the number of non-detects plus 1.	Advantage: • Provides an estimate of the value that is exceeded by 100(1-p)% of the underlying population Disadvantage: • Cannot be computed when the assumption on k is not valid

Box 2.4. Examples of Computing the Median, Trimmed Mean, and Winsorized Mean and Standard Deviation Using a Data Set that Contains Non-detects

The following examples use this data set of 12 measurements (after being ordered from smallest to largest): <0.15, <0.15, <0.15, 0.18, 0.25, 0.26, 0.27, 0.36, 0.50, 0.62, 0.63, 0.79 . Note three non-detects are in this data set, but each one has the same detection limit, 0.15. If multiple detection limits are present, consult a statistician for the best way to summarize the data.

Median

The median of the data set is (0.26 + 0.27) / 2 = 0.265. Note the non-detects do not have any impact on computing the median because fewer than half of the data were non-detects.

100p% Trimmed Mean

The percentage of non-detect measurements is 100(3/12) = 25%. Therefore, we set p = 0.25 and compute the 25% trimmed mean. (25% of n is 3.) Discard the smallest 0.25(12) = 3 and largest 3 measurements, that is, discard the three non-detects and the measurements 0.62, 0.63, 0.79. Compute the arithmetic mean on the remaining 6 measurements: Trimmed Mean = (0.18 + 0.25 + 0.26 + 0.27 + 0.36 + 0.50) / 6 = 0.30. This estimate is valid, if the underlying distribution of measurements is symmetric. If the distribution is not symmetric, this trimmed mean is a biased estimate.

Winsorized Mean

Replace the 3 non-detects by the next largest detected datum, which is 0.18. Replace the 3 largest measurements by the next smallest measurement, which is 0.50. Compute the arithmetic mean of the new set of 12 data: 0.18, 0.18, 0.18, 0.18, 0.25, 0.26, 0.27, 0.36, 0.50, 0.50, 0.50, 0.50.

$\overline{x}_w$ = (0.18 + 0.18 + 0.18 + 0.18 + 0.25 + 0.26 + 0.27 + 0.36 + 0.50 + 0.50 + 0.50 + 0.50) / 12 = 0.32 .

This estimate is valid if the underlying distribution of measurements is symmetric. If the distribution is not symmetric, this Winsorized mean is a biased estimate.

Winsorized Standard Deviation

Replace the 3 non-detects by the next largest detected datum, which is 0.18. Replace the 3 largest measurements by the next smallest measurement, which is 0.50. Compute the standard deviation, s, of the new set of 12 data:

$$s = [(0.18 - 0.32)^2 + (0.18 - 0.32)^2 + (0.18 - 0.32)^2 + (0.18 - 0.32)^2 + (0.25 - 0.32)^2 + (0.26 - 0.32)^2 + (0.27 - 0.32)^2 + (0.36 - 0.32)^2 + (0.50 - 0.32)^2 + (0.50 - 0.32)^2 + (0.50 - 0.32)^2 + (0.50 - 0.32)^2] / 11 = 0.1416$$

Compute $v = n - 2n' = 12 - 2(3) = 6$

Compute the Winsorized Standard Deviation:

$s_w = [s(n-1)]/(v-1) = [0.1416(11)] / 5 = 0.31$

This estimate is valid if the underlying distribution of measurements is symmetric. If the distribution is not symmetric, this Winsorized standard deviation is a biased estimate.

Box 2.5. Cohen Method for Computing the Mean and Variance of a Censored Data Set (EPA 1996; EPA 1998; Gilbert 1987, page 182)

- Let the single detection limit be denoted by DL. Let $x_1, x_2, \ldots, x_n$ denote the n measurements in the data set, including those that are less than DL. Let k be the number out of n that are greater than the DL.
- Compute h = (n-k)/n, which is the fraction of the n measurements that are below the DL.
- Compute the arithmetic mean of the k measurements that exceed the DL as follows

  $\overline{x}_c = (x_1 + x_2 + \ldots + x_k) / k$,

  where $x_1, x_2, \ldots,$ and x_k are all the measurements > DL.
- Compute the following statistic using the k measurements that exceed the DL:

  $s_c^2 = [(x_1 - \overline{x}_c)^2 + (x_2 - \overline{x}_c)^2 + \ldots + (x_k - \overline{x}_c)^2] / k$
- Compute $G = s_c^2 / (\overline{x}_c - DL)^2$
- Obtain the value of λ from Table A.5 for values of h and γ. Use linear interpolation in the table if necessary.
- Compute the Cohen mean and variance as follows:

  Cohen Mean $= \overline{x}_c - \lambda (\overline{x}_c - DL)$

  Cohen Variance $= s_c^2 + \lambda (\overline{x}_c - DL)^2$
- Cohen Standard Deviation is the square root of Cohen Variance.

Example:

- n = 25 measurements of a chemical in soil were obtained. One detection limit was equal to 36. Five measurements were reported as <36 (ND). The data obtained were:

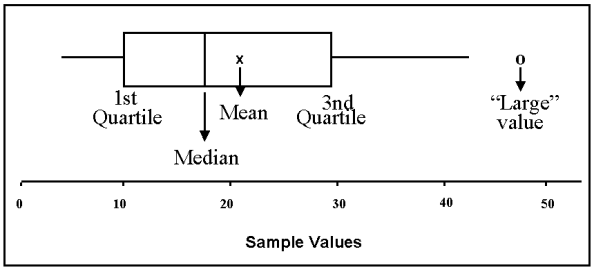
  <36, <36, <36, <36, <36, 49, 49, 59, 61, 62, 62, 65, 65, 65, 70, 72, 80, 80, 99, 99, 104, 110, 140 142, 144
- Compute h = (25 – 20)/25 = 0.20 = fraction of the 25 measurements that are below the detection limit
- Compute the arithmetic mean of the 20 measurements that exceed the detection limit:

  $\overline{x}_c = (49 + 49 + 59 + \ldots + 142 + 144) = 83.85$
- Compute $s_c^2 = [(49 - 83.85)^2 + (49 - 83.85)^2 + (59 - 83.85)^2 + \ldots + (142 - 83.85)^2 + (144 - 83.85)^2] / 20$
  $= 882.63$
- Compute $G = 882.63 / (83.85 - 36)^2 = 0.385$
- From Table A.5, we find by linear interpolation between $\gamma = 0.35$ and $\gamma = 0.40$ for h = 0.20 that $\lambda = 0.291$.
- Therefore, Cohen mean and variance are:

  Cohen Mean $= 83.85 - 0.291(83.85 - 36) = 69.9$

  Cohen Variance $= 882.63 + 0.291(83.85 - 36)^2 = 1548.9$
- Cohen Standard Deviation $= (1548.9)^{1/2} = 39.4$

2.4 Determining Presence of Data Outliers



Is the largest value an *outlier*. If so, should I delete it from the data set?

As discussed in Section 2.3, a set of data should always be carefully examined to determine the *center* of the data set and the spread or range of the data values. The center is usually characterized by computing the arithmetic mean, denoted by $\overline{x}$, and the *spread* by the standard deviation, s. In addition, look to see if any data seem much larger in value than most of the data. These unusually large data may be due to an error or they might indicate that small areas of much higher contamination levels are present at the site. Statistical tests for determining COPC (provided in Chapter 3) should not be conducted if the site or background data sets contain values that are mistakes that occurred during the collection, handling, measurement, and documentation of samples. If some of the data are so large as to cause concern that a mistake has been made, a statistical test for outliers should be conducted. If the test indicates the suspect value(s) are indeed larger than expected, relative to the remaining data, the outliers should be examined to determine if they are mistakes or errors. If they are, they should be removed from the data set. Otherwise, they should not be removed, even though the statistical test indicated they were outliers.

The general rule is that a measurement should never be deleted from a data set *solely* on the basis of an outlier test. This rule is used because outlier tests compare suspect data points with what is believed to be the true underlying distribution of the data, for example, a normal or lognormal distribution. Hence, the outlier test may give the wrong answer because the assumed underlying distribution is not the correct choice. Suppose, for example, that the underlying distribution was assumed to be normal for purposes of conducting an outlier test, but in fact the underlying distribution was lognormal. In that case, a suspect large value could be incorrectly identified as an outlier by the test because such large values are not consistent with the underlying assumption of a normal distribution.

For all outlier tests discussed, except the Walsh test, a test for normality should be performed on the data set. The normality test is conducted on the data set after the suspected outlier(s) is deleted. The following tests for outliers are described and illustrated in Box 2.6: the Dixon test, the Discordance test, the Rosner test, and the Walsh test . The assumptions, advantages, and disadvantages of each test are provided in Box 2.6. The first three tests are described and illustrated in EPA (1996). The discussion of the Walsh test is from EPA (1998), which corrects some errors that occurred in the description of that test in EPA (1996).

Box 2.6. Assumptions, Advantages, and Disadvantages of Outlier Tests

Statistical Test	Assumptions	Advantages/Disadvantages
Dixon Test	• $n \leq 25$ • Measurements are representative of the underlying population. • The measurements without the suspect outlier are normally distributed; otherwise, see a statistician. • Test can be used to test for either one suspect large outlier or one suspect small outlier. The latter case is not considered here as it is not of interest for determining a COPC.	Advantages: • Simple to compute by hand • The test is available in the DataQUEST software EPA (1997). Disadvantage: • Test should be used for only one suspected outlier. Use the Rosner test if multiple suspected outliers are present. • Must conduct a test for normality on the data set after deleting the suspect outlier and before using the Dixon test
Discordance Test	• $3 < n \leq 50$ • Measurements are representative of underlying population. • The measurements without the suspected outlier are normally distributed; otherwise, see a statistician. • Test can be used to test that the largest measurement, if a suspected outlier or the smallest measurement is a suspected outlier. The latter case is not considered here as it is not of interest for determining COPC.	Advantages: • Simple to compute by hand • The test is available in the DataQUEST software EPA (1997). Disadvantages: • Test can be used for only one suspected outlier. Use the Rosner test if there are multiple suspected outliers. • Must conduct a test for normality on the data set after deleting the suspect outlier and before using the Discordance test.
Rosner's Test	• $n \geq 25$ • Measurements are representative of underlying population. • The measurements without the suspected outliers are normally distributed; otherwise, see a statistician.	Advantages: • Can test for up to 10 outliers • The test is available in the DataQUEST software EPA (1997). Disadvantages: • Must conduct a test for normality after deleting the suspected outliers and before using Rosner's test • Computations are more complex than for Dixon's Test or the Discordance Test
Walsh's Test	• $n > 60$ • Measurements are representative of the underlying population. • Test can be used to test that the largest r measurements or the smallest r measurements are suspected outliers. The latter case (discussed in EPA 1998) is not considered here as it is not of interest for determining COPC.	Advantages: • Can test for 1 or more outliers • The measurements need not be normally distributed. • Need not conduct a test for normality before using the test • The test is available in the DataQUEST software EPA (1997). Disadvantages: • Must have $n > 60$ to conduct the test • The test can only be performed for the $\alpha = 0.05$ and 0.10 significance levels, and the α level used depends on n: the $\sigma = 0.05$ level can only be used if $n > 220$ and the $\sigma = 0.10$ level can only be used if $60 < n \leq 220$. • Test calculations are more complex than for the Dixon test or the Discordance test.

		• The number of identified suspected outliers, r, are accepted or rejected as a group rather than one at a time

The procedures for conducting the Dixon Extreme Value Test, the Discordance Test, and the Walsh Test, with an example for each, are provided in Box 2.7, Box 2.8, and Box 2.9, respectively. The Rosner test is described in Box 2.10 and illustrated in Box 2.11.

Before preceding further, we ask "What is a statistical test?" A statistical test is a comparison of some data-based quantity (test statistic) with a critical value that is usually obtained from a special table. This comparison (test) is conducted to determine if a statistically significant result has occurred. The statistical test is evaluating whether the data obtained (as summarized in the test statistic) are convincing beyond a reasonable doubt that a specified null hypothesis, denoted by H_o, is false and should be rejected in favor of a specified alternative hypothesis, H_a, that is true and should be accepted. In this handbook the following H_o and H_a are used when testing for outliers:

H_o: The suspect (unusually large) data are from the same underlying probability distribution as the other data in the data set.

H_a: The suspect data <u>are not</u> from the same underlying probability distribution as the other data in the data set.

If the test rejects the H_o in favor of the H_a, then we can conclude with $100(1-\alpha)\%$ confidence the suspect data really are outliers and hence should be examined closely to see if they are due to errors or if they are an indication of the presence of areas where concentrations are higher than for most of the site. If the test does not reject H_o, either the suspect data are really from the same distribution as the remaining data, or the information in the data set is simply not sufficient for the test to reject H_o with the required confidence.

The quantity α is a value less than 0.50 and greater than zero. α is the probability we can tolerate of falsely rejecting H_o and accepting H_a, that is, of falsely concluding the suspect data are outliers.

Box 2.7. The Dixon Extreme Value Test (EPA 1998)

- Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the n measurements in the data set after they have been listed in order from smallest to largest. The parentheses around the subscripts indicates the measurements are ordered from smallest to largest.
- $x_{(n)}$ (the largest measurement) is suspected of being an outlier.
- Perform test for normality on $x_{(1)}$ through $x_{(n-1)}$.
- Specify the tolerable decision error rate, α (significance level), desired for the test. α may only be set equal to 0.01, 0.05 or 0.10 for the Dixon test.
- Compute

$$
\begin{aligned}
C &= [x_{(n)} - x_{(n-1)}] / [x_{(n)} - x_{(1)}] \quad \text{if} \quad 3 \le n \le 7 \\
&= [x_{(n)} - x_{(n-1)}] / [x_{(n)} - x_{(2)}] \quad \text{if} \quad 8 \le n \le 10 \\
&= [x_{(n)} - x_{(n-2)}] / [x_{(n)} - x_{(2)}] \quad \text{if} \quad 11 \le n \le 13 \\
&= [x_{(n)} - x_{(n-2)}] / [x_{(n)} - x_{(3)}] \quad \text{if} \quad 14 \le n \le 25
\end{aligned}
$$

- If C exceeds the critical value in Table A.2 for the specified n and α, then declare that $x_{(n)}$ is an outlier and should be investigated further.

Example: Suppose the ordered data set is 34, 50, 52, 60, 62. Suppose we wish to test if 62 is an outlier from an assumed normal distribution for the n = 5 data. Perform a test for normality on the data 34, 50, 52, 60. We note that any test for normality will have little ability to detect non-normality on the basis of only 4 data values. (See Section 2.6 for statistical methods of testing the normality assumption.) Suppose α is selected to be 0.05, that is, we want no more than a 5% chance the test will incorrectly declare the largest observed measurement is an outlier. Compute C = (62 - 60)/(62 – 34) = 0.071. Determine the test critical value from Table A.2. The critical value is 0.642 when n = 5 and α = 0.05. As 0.071 is less than 0.642, the data do not indicate the measurement 62 is an outlier from an assumed normally distribution.

Box 2.8. Discordance Outlier Test (EPA 1998)

- Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the n measurements in the data set after they have been listed in order from smallest to largest.
- $x_{(n)}$ (the largest measurement) is suspected of being an outlier.
- Specify the tolerable decision error rate, α (significance level) desired for the test. α may be specified to be 0.01 or 0.05 for the Discordance Outlier test.
- Compute the sample arithmetic mean, $\bar{x}$, and the sample standard deviation, s.
- Compute $D = [x_{(n)} - \bar{x}\ ] / s$
- If D exceeds the critical value from Table A.3 for the specified n and α, $x_{(n)}$ is an outlier and should be further investigated.

Example: Suppose the ordered data set is 34, 50, 52, 60, 62. We wish to test if 62 is an outlier from an assumed normal distribution for the data. Suppose α is selected to be 0.05. Using the n = 5 data, we compute $\bar{x}$ = 51.6 and s = 11.08. Hence, D = (62 – 51.6) / 11.08 = 0.939. The critical value from Table A.3 for n = 5 and α = 0.05 is 1.672. As 0.939 is less than 1.672, the data do not indicate the measurement 62 is an outlier from an assumed normally distribution.

Box 2.9. The Walsh Outlier Test (EPA 1998)

- Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote n measurements in the data set after they have been listed in order from smallest to largest. Do not apply the test if $n < 60$. If $60 < n \leq 220$, then use $\alpha = 0.10$. If $n > 220$, then use $\alpha = 0.05$.
- Identify the number of possible outliers, r, where r can equal 1.
- Compute: $c = [(2n)^{1/2}]$, $k = r + c$, $b^2 = 1/\alpha$,
  $a = (1 + b\{(c-b^2)/(c-1)\}^{1/2}) / (c - b^2 - 1)$
  where [ ] indicates rounding the value to the largest possible integer (that is, 3.24 becomes 4).
- The Walsh test declares that the r largest measurements are outliers (with a α level of significance) if
  $$x_{(n+1-r)} - (1 + a)x_{(n-r)} + a x_{(n+1-k)} > 0$$

Example: Suppose n = 70 and that r = 3 largest measurements are suspected outliers. The significance level α = 0.10 must be used because $60 < n \leq 220$. That is, we must accept a probability of 0.10 the test will incorrectly declare that the 3 largest measurements are outliers.

- Compute $c = [(2x70)^{1/2}] = 12$, $k = 3 + 12 = 15$, $b^2 = 1 / 0.10 = 10$,
  $a = \{1 + 3.162\{(12 – 10) / (12 – 1)\}^{1/2}\} / (12 – 10 – 1) = 2.348$
- $x_{(n+1-r)} = x_{(70+1-3)} = x_{(68)}$ is the 68^{th} largest measurement (two measurements are larger)
  $x_{(n-r)} = x_{(70-3)} = x_{(67)}$ is the 67^{th} largest measurement
  $x_{(n+1-k)} = x_{(70+1-15)} = x_{(56)}$ is the 56^{th} largest measurement
- Order the 70 measurements from smallest to largest. Suppose $x_{(68)} = 83$, $X_{(67)} = 81$, and $x_{(56)} = 20$.
- Compute $x_{(n+1-r)} - (1+a)x_{(n-r)} + a x_{(n+1-k)} = 83 – (1+2.348)81 + 2.348(20) = -141.22$ which is smaller than 0. Hence, the Walsh test indicates that the r = 3 largest measurements are not outliers.

Box 2.10. The Rosner Outlier Test (EPA 1996)

STEP 1:

- Select the desired significance level α, that is, the probability that can be tolerated of the Rosner test falsely declaring that outliers are present.
- Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote n measurements in the data set after they have been listed in order from smallest to largest, where $n \geq 25$.
- Identify the maximum number of possible outliers, denoted by r.

STEP 2:

- Set i = 0 and use the following formulas

$$\bar{x}^{(i)} = (x_1 + x_2 + \ldots + x_{n-i}) / (n - i)$$

$$s^{(i)} = \{[(x_1 - \bar{x}^{(i)})^2 + (x_2 - \bar{x}^{(i)})^2 + \ldots + (x_{n-i} - \bar{x}^{(i)})^2] / (n - i)\}^{1/2}$$

  to compute the sample arithmetic mean, labeled $\bar{x}^{(0)}$, and $s^{(0)}$ using all n measurements. Determine the measurement that is farthest from $\bar{x}^{(0)}$ and label it $y^{(0)}$
- Delete $y^{(0)}$ from the data set of n measurements and compute (using i = 1 in the above formulas) the sample arithmetic mean, labeled $\bar{x}^{(1)}$, and $s^{(1)}$ on the remaining n-1 measurements. Determine the measurement that is farthest from $\bar{x}^{(1)}$ and label it $y^{(1)}$.
- Delete $y^{(1)}$ from the data set and compute (using i = 2 in the above formulas) the sample arithmetic mean, labeled $\bar{x}^{(2)}$, and $s^{(2)}$ on the remaining n-2 measurements.
- Continue using this process until the r largest measurements have been deleted from the data set.
- The values of $\bar{x}^{(0)}, \bar{x}^{(1)}, \ldots, s^{(0)}, s^{(1)}, \ldots$ are computed using the following formulas:

STEP 3:

- To test if there are r outliers in the data set compute

$$R_r = [\,|y^{(r-1)} - \bar{x}^{(r-1)}|\,] / s^{(r-1)}$$

- Determine the critical value λ_r from Table A.4 for the values of n, r, and α.
- If R_r exceeds λ_r, conclude r outliers are in the data set.
- If not, test if r-1 outliers are present. Compute
  $R_{r-1} = [\,|y^{(r-2)} - \bar{x}^{(r-2)}|\,] / s^{(r-2)}$
- Determine the critical value λ_{r-1} from Table A-4 for the values of n, r - 1 and α.
- If R_{r-1} exceeds λ_{r-1}, conclude r –1 outliers are in the data set.
- Continue on in this way until either it is determined that there are a certain number of outliers are present or that no outliers exist at all.

Box 2.11. Example: Rosner Outlier Test

STEP 1:

Consider the following 32 data points (in ppm) listed in order from smallest to largest: 2.07, 40.55, 84.15, 88.41, 98.84, 100.54, 115.37, 121.19, 122.08, 125.84, 129.47, 131.90, 149.06, 163.89, 166.77, 171.91, 178.23, 181.64, 185.47, 187.64, 193.73, 199.74, 209.43, 213.29, 223.14, 225.12, 232.72, 233.21, 239.97, 251.12, 275.36, and 395.67.

A normal probability plot of the data identified four potential outliers: 2.07, 40.55, 275.36 and 395.67. Moreover, a normal probability plot of the data set after excluding the four suspect outliers provided no evidence that the data are not normally distributed.

STEP 2:

First use the formulas in Box 2.10 to compute $\overline{x}^{(0)}$ and $s^{(0)}$ using the entire data set. Using subtraction, it was found that 395.67 was the farthest data point from $\overline{x}^{(0)}$, so $y^{(0)} = 395.67$. Then 395.67 was deleted from the data set and $\overline{x}^{(1)}$ and $s^{(1)}$ are computed on the remaining data. Using subtraction, it was found that 2.07 was the farthest value from $\overline{x}^{(1)}$, so $y^{(1)} = 2.07$. This value was then dropped from the data and the process was repeated to determine $\overline{x}^{(2)}$, $s^{(2)}$, $y^{(2)}$ and $\overline{x}^{(3)}$, $s^{(3)}$, $y^{(3)}$. These values are summarized below:

i	$\overline{x}^{(i)}$	$s^{(i)}$	$y^{(i)}$
0	169.92	73.95	395.67
1	162.64	62.83	2.07
2	167.99	56.49	40.55
3	172.39	52.18	275.36

STEP 3:

To apply the Rosner test, first test if 4 outliers are present. Compute

$$R_4 = |y^{(3)} - \overline{x}^{(3)}| / s^{(3)} = |275.36 - 172.39| / 52.18 = 1.97$$

Suppose we want to conduct the test at the $\alpha = 0.05$ level, that is, we can tolerate a 5% chance of the Rosner test falsely declaring 4 outliers. In Table A.4, we find $\lambda_4 = 2.89$ when $n = 32$, $r = 4$ and $\alpha = 0.05$. As $R_4 = 1.97$ is less than 2.89, we conclude that 4 outliers are not present. Therefore, test if 3 outliers are present. Compute

$$R_3 = |y^{(2)} - \overline{x}^{(2)}| / s^{(2)} = |40.55 - 167.99| / 56.49 = 2.26$$

In Table A.4 we find $\lambda_3 = 2.91$ when $n = 32$, $r = 3$ and $\alpha = 0.05$. Because $R_4 = 2.26$ is less than 2.91, we conclude that 3 outliers are not present. Therefore, test if 2 outliers are present. Compute

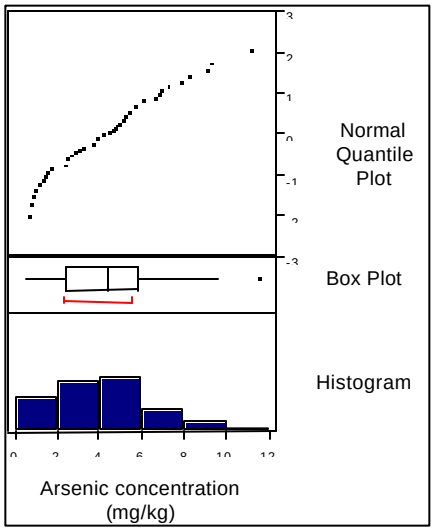
$$R_2 = |y^{(1)} - \overline{x}^{(1)}| / s^{(1)} = |2.07 - 162.64| / 62.83 = 2.56$$

In Table A.4, we find $\lambda_2 = 2.92$ for $n = 32$, $r = 2$ and $\alpha = 0.05$. As $R_2 = 2.56$ is less than 2.92, we conclude that 2 outliers are not present in the data set. Therefore, test if 1 outlier is present. Compute

$$R_1 = |y^{(0)} - \overline{x}^{(0)}| / s^{(0)} = |395.67 - 169.92| / 73.95 = 3.05$$

In Table A-4 we find $\lambda_1 = 2.94$ for n = 32, r = 1and $\alpha = 0.05$. Since $R_1 = 3.05$ is greater than 2.94, we conclude at the $\alpha = 0.05$ significance level that 1 outlier is present in the data set. Therefore, the measurement 395.67 is considered to be a statistical outlier. It will be further investigated to determine if it is an error or a valid data value.

2.5 Graphical Data Analysis



Different views of the data tell different stories.

Graphical plots of the site and background data sets are extremely useful and necessary tools to:

- conduct exploratory data analyses to develop hypotheses about possible differences in the means, variances, and shapes for the site and background measurement distributions
- visually depict and communicate differences in the distribution parameters (means, variances, and shapes) for the site and background measurement distributions
- graphically evaluate if the site and background data have a normal, lognormal, or some other distribution
- evaluate, illuminate, and communicate the results obtained using formal statistical tests for COPC (Section 3.0).

In this section, we discuss and illustrate four types of graphical plots: histograms, boxplots, quantile plots and probability plots. Much of this discussion is from EPA (1996), which offers a more thorough survey of graphical methods, including plots for two or more variables and for data collected over time and space. The four methods included in this handbook, summarized in Box 2.12, were selected because they are quite simple and well suited for use with formal statistical tests (Section 3.0) to distinguish between site and background data sets, that is, to identifying COPC. The methods in Box 2.12 can be easily generated using the DataQUEST (EPA 1997) statistical software.

Box 2.12. Summary of Selected Graphical Methods and Their Advantages and Disadvantages

Method	Description	Advantages	Disadvantages
Histogram	A graph constructed using bars that describes the approximate shape of the data distribution	• Easy to construct, understand and explain • Shows the distribution shape, spread (range), and central tendency (location)	• The choice of interval width for the histogram bars can affects the perception of the shape of the distribution • The histogram can be misleading unless the number of measurements used in its construction is reported on the graph.
Boxplot	A simple box with extended lines (*whiskers*) that depict the central tendency and shape of the distribution.	• Easy to construct, understand and explain • Shows the 25^{th} , 50^{th} and 75^{th} percentiles as well as the mean, spread of the data, and extreme values • Good for comparing multiple, for example, site and background, data sets on a common scale on the same page of report	• Provides less detailed information about the shape of the distribution than is conveyed by the histogram
Quantile Plot	A plot of each data value versus the fraction of the data set that is less than that value	• Easy to construct • No assumption is made about the shape of the data distribution • The quantiles (or percentiles) of the data set can be read from the plot • The plot indicates if the distribution of the data set is symmetric or asymmetric	• Somewhat more difficult to understand and explain than histograms or boxplots • Must know how to interpret the shape of plot to determine if the data set is symmetric or asymmetric .• Interpretation of the shape of the plot line is subjective • Not as effective as a probability plot for evaluating the distribution model (for example, normal or lognormal)

Probability Plot	A plot of the estimated quantiles of a data set versus the quantiles of a hypothesized distribution for the data set.	• A subjective, graphical method for testing whether a data set may be well fit by an hypothesized distribution, such as the normal or lognormal • Deviations of the plotted points from a straight line provides information about how the data set deviates from the hypothesized distribution. • Provides initial hints about whether the data set might be composed of two or more distinct populations, for example, background and site contamination populations	• A separate plot is required for each hypothesized distribution. • The plots require either special probability plotting paper or a table for determining the quantiles of the hypothesized distribution. • Subjective judgment is used to decide if the plot indicates the data set may have the same distribution as the hypothesized distribution. • The plot should be used in conjunction with a formal test for distribution (Section 2.6). • Hints about possible multiple populations (for example, differences in site and background) *must* be confirmed by formal statistical tests (Section 3.0), histograms, and geochemical analyses and expert judgment (Section3.1).

2.5.1 Histograms

The histogram is a bar chart of the data that displays how the data are distributed over the range of measured values on the x-axis. The general shape of the histogram is used to assess whether a large portion of the data is tightly clustered around a central value (the mean or median) or spread out over a larger range of measured values. If the histogram has a symmetric shape, it suggests the underlying population might be normally distributed, whereas an asymmetric shape with a long tail of high measurement values may suggest a lognormal or some other skewed distribution. These hypotheses can be evaluated using probability plots (Section 2.5.4) and the methods in Section 2.6. Figure 2.5-1 is a histogram of 22 measurements.

The histogram is constructed by first dividing the range of measured values into intervals. The number of measurements within each interval is counted and this count is divided by the

total number of measurements in the data set to obtain a percentage. The length of the bar for that interval is the magnitude of the computed percentage. The sum of the bar percentages is 100%. Directions for constructing a histogram are provided in Box 2.13. An example is provided in Box 2.14.

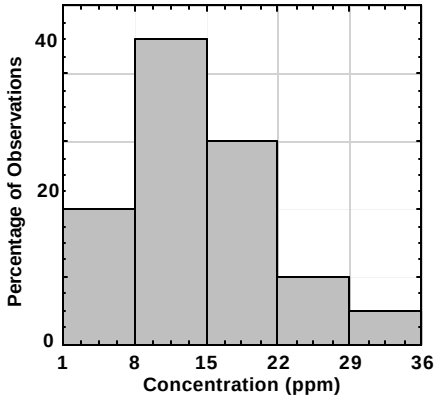


Figure 2.5-1. Histogram Example

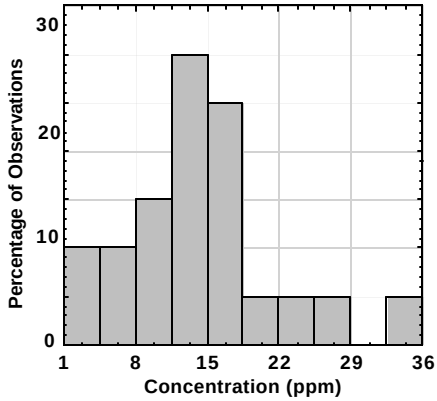


Figure 2.5-2. Histogram with Smaller Interval Widths

The visual impression conveyed by a histogram is quite sensitive to the choice of the interval (width of the bar). The histogram in Figure 2.5-2 is based on the same data as that used for Figure 2.5-1, but it uses an interval (bar width) of 3.5 ppm rather than 7 ppm. Note that Figure 2.5-2 gives the impression the data distribution is more skewed to the right (toward larger values) than does Figure 2.5-1. That impression is only due to the use of a smaller interval. Only 3 data values are greater than 22 ppm, so the amount of information available to define the shape and extent of the right tail of the distribution is very limited. To guard against misinterpretations of histograms, the number of data points used to construct the histogram must always be reported. The bar widths should not be too narrow if the number of data is small. It is useful to construct histograms for two or three different bar widths and select the bar width that provides the most accurate picture of the data set. All interval widths in a histogram should be the same size, as is the case for Figures 2.5-1 and 2.5-2.

Box 2.13. Directions for Constructing a Histogram (after EPA 1996, page 2.3-2)

STEP 1: Let x_1 , x_2 , ..., x_n represent the n measurements. Select the number of intervals (bar widths), each of equal width*. A rule of thumb is to have between 7 and 11 intervals that cover the range of the data. Specify a rule for deciding which interval a data point is assigned to, if a measurement should happen to equal in value an interval endpoint.

STEP 2: Count the number of measurements within each interval.

STEP 3: Divide the number of measurements within each interval by n (the total number of measurements in the data set) to compute the percentage of measurements in each interval.

STEP 4: For each interval, construct a box whose length is the percentage value computed in Step 3.

* EPA (1996, Box 2.3-2) considers the case where the bar widths are not of equal size.

Box 2.14. Example: Constructing a Histogram (from EPA 1996, page 2.3-2).

STEP 1: Suppose the following n = 22 measurements (in ppm) of a chemical in soil have been obtained:

17.7, 17.4, 22.8, 35.5, 28.6, 17.2 19.1, <4, 7.2, <4, 15.2, 14.7, 14.9, 10.9, 12.4, 12.4, 11.6, 14.7, 10.2, 5.2, 16.5, and 8.9.

These data range from <4 to 35.5 ppm. Suppose equal sized interval widths of 5 ppm are used, that is, 0 to 5, 5 to 10, 10 to 15, etc. Also, suppose we adopt the rule that a measurement that falls on an interval endpoint will be assigned to the highest interval containing the value. For example, a measurement of 5 ppm will be placed in the interval 5 to 10 ppm instead of 0 to 5 ppm. For this particular data set, no measurements happen to fall on 5, 10, 15, 20, 25, 30, or 35. Hence, the rule is not needed for this data set.

STEP 2: The following table shows the number of observations within each interval defined in Step 1.

STEP 3: The table contains n = 22 measurements, so the number of observations in each interval will be divided by 22. The resulting percentages are shown in column 3 of the table.

STEP 4: For the first interval (0 to 5 ppm), the vertical height of the bar is 9.10. For the second interval (5 to 10 ppm), the height of the bar is 13.6, and so forth for the other intervals.

Interval	Number of Data in Interval	Percent of Data in Interval
0 - 5 ppm	2	9.10
5 - 10 ppm	3	13.60
10 - 15 ppm	8	36.36
15 - 20 ppm	6	27.27
20 - 25 ppm	1	4.55
25 - 30 ppm	1	4.55
30 - 35 ppm	0	0.00
35 - 40 ppm	1	4.55

2.5.2 Boxplots

The boxplot, sometimes called a *box-and-whisker* plot, simultaneously displays the full range of the data, as well as key summary statistics. Figure 2.5-3 shows a boxplot of the data listed in Step 1 of Box 2.14. (In this plot, the two <4 values were set equal to 4.)

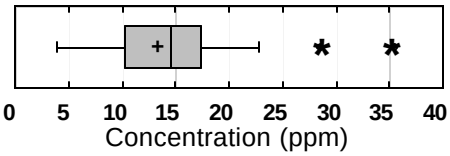


Figure 2.5-3. Example: Boxplot (Box-and-Whisker Plot).

The boxplot is composed of a central box divided by a vertical line (that is placed at the median value of the data set) and two lines extending out from the box (called the whiskers). The length of the central box (the interquartile range; see Box 2.1 for definition) indicates the

spread of the central 50% of the data, while the lengths of the whiskers show the extent that measurements are spread out below and above the central 50% box. The upper end of the whisker that extends to higher concentrations is the largest data value that is less than the 75th percentile plus 1.5 times the length of the 50% box. Similarly, the lower end of the whisker that extends to lower concentrations is the smallest data value that is greater than the 25th percentile minus 1.5 times the length of the 50% box. As previously noted, the median of the data set is displayed as a vertical line through the box. The arithmetic mean of the data set is displayed using a + sign. Any data values that fall outside the range of the whiskers are displayed by an *. The boxplot as shown in Figure 2.5-3 is sometimes rotated 90 degrees counter-clock-wise, so that the whickers are vertical rather than horizontal.

The boxplot provides a visual picture of the symmetry or asymmetry of the data set. If the data set distribution is symmetric, the central box will be divided into two equal halves by the median, the mean will be approximately equal to the median, the whiskers will be approximately the same length, and approximately the same number of extreme data points (if any exist) will occur at either end of the plot. EPA (1996, page 2.3-5) illustrates how to construct a boxplot.

2.5.3 Quantile Plots

The quantile plot shows each data value plotted versus the fraction (f) of the entire data set that is less than that value. The plot derives its name from the fact that the quantiles of the data set can be read directly from the y-axis of the plot. A quantile is the same as a percentile (defined in Box 2.1) except that it is expressed as a fraction rather than a percentage. Figure 2.5-4 shows an example of a quantile plot.

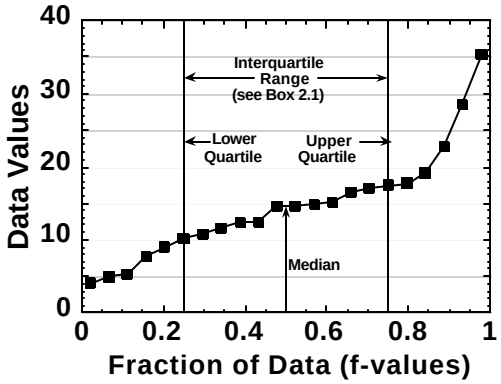


Figure 2.5-4. Example: Quantile Plot of a Skewed Data Set

The y (vertical) axis of a quantile plot shows the value of each individual data point. The x (horizontal) axis ranges from 0.0 to 1.0. The plotted data values that fall between vertical lines drawn at fractions 0.25 and 0.75 are those that are within the central 50% box of a boxplot. The difference between the data value at the 0.75 fraction and the data value at the 0.25 fraction is the interquartile range of the data set (Box 2.1). In Figure 2.5-4, it appears that the 0.75 quantile (75th percentile) is approximately 17.5 (on the y-axis) and the 0.25 quantile (25th percentile) is about 10.0. Hence, the interquartile range is about 17.5 – 10.0 =

7.5. From the plot, the 0.50 quantile (50^{th} percentile or median data value) appears to be about 15.0.

The shape of the plotted points on the quantile plot can be used to assess whether the data set is symmetric or skewed. The plotted curve for a data set that is skewed to the *right* has a steeper slope at the top right than at the bottom left, as in Figure 2.5-4. The plotted curve for a data set that is skewed to the *left* has a steeper slope near the bottom left of the graph. If the data set has a symmetric shape, the top portion of the graph will stretch to the upper right corner in the same way the bottom portion of the graph stretches to the lower left, creating an S-shape curve.

Box 2.15 provides directions for generating a quantile plot. An example is provided in Box 2.16.

Box 2.15. Directions for Constructing a Quantile Plot

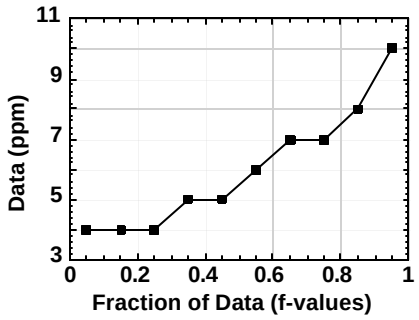
Let $x_1, x_2, ..., x_n$ represent the n data points. Let $x_{(i)}$, for i = 1, 2, …, n, be the data listed in order from smallest to largest, so that $x_{(1)}$ is the smallest, $x_{(2)}$ is the second smallest, … , and $x_{(n)}$ is the largest. For each i, compute the fraction $f_i = (i - 0.5)/n$. The quantile plot is a plot of the n pairs ($x_{(i)}$, f_i), with straight lines connecting the plotted points.

Box 2.16. Example: Constructing a Quantile Plot

Consider the following 10 data points (in ppm): 4, 5, 6, 7, 4, 10, 4, 5, 7, and 8. The data ordered from smallest to largest, $x_{(1)}$, $x_{(2)}$, …., $x_{(n)}$, are shown in the first column of the table below and the ordered number for each observation, i, is shown in the second column. The third column displays the values f_i for each i where $f_i = (i - 0.5)/n$.

$x_{(i)}$	i	f_i	$x_{(i)}$	i	f_i
4	1	0.05	6	6	0.55
4	2	0.15	7	7	0.65
4	3	0.25	7	8	0.75
5	4	0.35	8	9	0.85
5	5	0.45	10	10	0.95

The pairs $(x_{(i)}, f_i)$ are then plotted to yield the following quantile plot:



2.5.4 Probability Plots

A probability plot is a graph of data plotted versus the quantiles of a user-specified distribution. Usually, the goal of constructing a probability plot is to visually (subjectively) evaluate the null hypothesis that the data are well fit (modeled) by the specified distribution. Frequently, the null hypothesis is the data set has a normal or lognormal distribution, although other distributions such as the Weibull and Gamma distributions (Gilbert 1987, page 157) are sometimes used. If the graph of plotted points in a probability plot appears linear to the eye with little scatter or deviation about the line, one would conclude the data appear to be well fit by the specified distribution. If the plotted points do not approximate a straight line, the type of departures from linearity provide information about how the actual data distribution deviates from the hypothesized distribution. Probability plots should always be used in conjunction with one of the formal statistical tests discussed in Section 2.6 for evaluating what the best fitting distribution may be for the data set may be.

Figure 2.5-5 shows a probability plot for some typical concentration data. The null hypothesis used to obtain this plot was the data follow the normal distribution. However, the plotted points are not well fit by a straight line. Hence, we should reject the null hypothesis that data are normally distributed.

The probability plot in Figure 2.5-5 was obtained by plotting each data value versus its expected quantile, assuming that the data are indeed distributed as a standard normal distribution. The expected quantiles for a standard normal distribution are obtained from Table A.1, as illustrated in Figure 2.5-5. We note that if the null hypothesis had been that the data follow a *lognormal* distribution, the *logarithm* of each datum would have been plotted versus its expected quantile. Now, because the logarithms of lognormally distributed data have a normal distribution, Table A.1 is used to obtain expected quantiles when evaluating the fit of a lognormal distribution. A statistician should be consulted to determine how to obtain expected quantiles for hypothesized distributions other than the normal or lognormal.

A special type of graph paper called probability-plotting paper can be used in order to avoid the need to determine the expected quantiles of the hypothesized distribution. If the hypothesized distribution is the normal distribution, the data values are plotted on normal (distribution) probability paper. If the hypothesized distribution is the lognormal distribution, the logarithms of the data are plotted on normal probability paper.

Figure 2.5-6 is a probability plot constructed using normal probability paper to test the null hypothesis that data have a normal distribution. The data set used for Figure 2.5-6 was also used to construct Figure 2.5-5, so the plots are identical. Note the x-axis for Figure 2.5-6 represents cumulative probabilities (rather than quantiles) for the standard normal distribution.

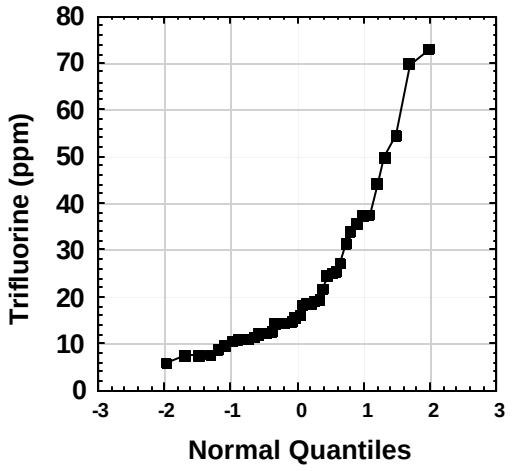


Figure 2.5-5. Example: Probability Plot for Which the Hypothesized Distribution is Normal (Quantiles on the x-Axis).

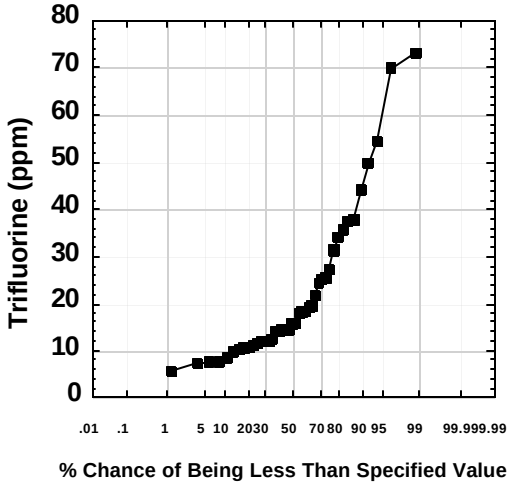


Figure 2.5-6. Example: Probability Plot for Normal Hypothesized Distribution (100 x Probability on the x-Axis).

Similar to quantile plots, the shape of probability plots provide information about the shape of the data distribution. If one constructs a probability plot assuming the data are normally distributed, but the data set is actually skewed to the right, the normal probability plot graph will be convex. If the data set is skewed to the left, the graph will be concave. The plotted points in Figures 2.5-5 and 2.5-6 form a convex curve, indicating the data set is skewed to the right. Because lognormal distributions are right-skewed, it is logical to test the hypothesis that the data set is well fit by a lognormal distribution. Figure 2.5-7 shows a probability plot of the *logarithms* of the data plotted versus quantiles of the normal distribution. As the

plotted line is well fit by a straight line, we may tentatively accept the hypothesis that data are lognormally distributed. However, this result should be checked by conducting the Shapiro-Wilk W test discussed and illustrated in Section 2.6.1.

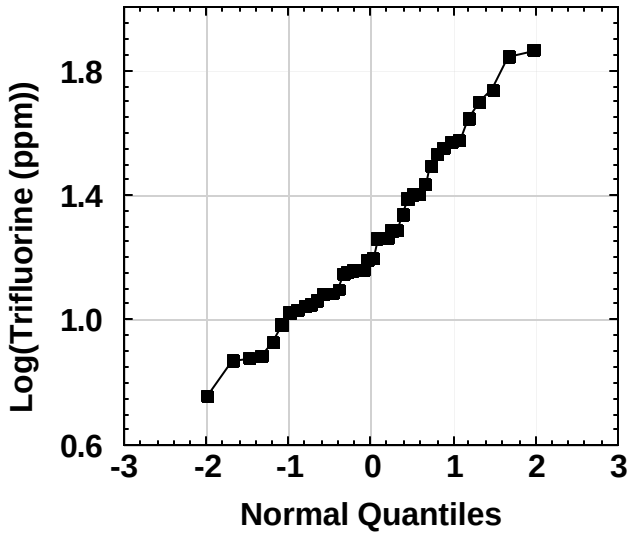


Figure 2.5-7. Example of a Probability Plot to Test that the Data have a Lognormal Distribution

Box 2.17 provides directions for constructing a probability plot when the null hypothesis is the data are normally distributed. The same procedure would be used to test the null hypothesis that data are lognormally distributed, except that logarithms of the data instead of the untransformed x_i data are used. Box 2.18 provides an example of this procedure. These directions provide a method for handling data sets for which some measurements occur more than once, that is, tied data values are present. Ties are managed by computing cumulative frequencies and plotting points on the plot only for distinct (different) data values. An alternative procedure is to plot a point for each measurement, whether or not it is the same value as another datum. This construction is done in the manner illustrated for constructing quantile plots (Box 2.16).

Box 2.17. Directions for Constructing a Normal Probability Plot (from EPA 1996, page 2.3-10)

Let $x_1, x_2, \ldots, x_n$ represent the n data points. We desire to test if the n data are normally distributed. We do so by constructing a normal probability plot.

STEP 1: Order all the n data from smallest to largest and denote the ordered *distinct* (different) data values by $x_{(1)}, x_{(2)}, \ldots, x_{(n')}$, where n' may be less than n. For each distinct data value, compute the absolute frequency, AF_i. The absolute frequency is the number of times each distinct value occurs. If a data value occurs only once, the absolute frequency for that value is 1. If a data value occurs more than once, count the number of times the distinct value occurs. For example, consider the data set 1, 2, 3, 3, for which n = 4 and $n' = 3$. The absolute frequency of value 1 is 1, that is, $AF_1 = 1$. The absolute frequency of value 2 is 1, that is, $AF_2 = 1$. But the absolute frequency of value 3 is 2, that is, $AF_3 = 2$, as 3 appears 2 times in the data set.

STEP 2: Compute the cumulative frequency, CF_i, for each of the n' distinct data values. The CF_i is the number of data points that are less than or equal to $x_{(i)}$, that is, $CF_i = \sum_{j=1}^{i} AF_j$. Using the data given in Step 1, the CF for value 1 is 1, the CF for value 2 is 2 (that is, 1+1), and the CF for value 3 is 4 (that is, 1+1+2).

STEP 3: Compute $Y_i = \dfrac{CF_i}{(n+1)}$ for each distinct data value

STEP 4: Determine from the standard normal distribution (Table A.1) the quantile associated with each value of Y_i. Denote the quantile of the i^{th} distinct data value by Q_i. We note the EPA DataQUEST software (EPA 1997) will construct probability plots, saving the effort of determining quantiles from special tables and plotting the points.

STEP 5: Plot the pairs (x_i, Q_i). If the plot of these points is well fit by a straight line, we may conclude the data are probably distributed normally. Otherwise, the data may be better fit by another distribution.

Box 2.18. Example: Constructing a Normal Probability Plot

Consider the following n = 15 data points that have been ordered from smallest to largest: 5, 5, 6, 6, 8, 8, 9, 10, 10, 10, 10, 10, 12, 14, and 15. We wish to test the data are normally distributed by constructing a normal probability plot.

STEP 1: The data set contains $n' = 8$ distinct data values. Because the value 5 appears 2 times, its absolute frequency is 2, or $AF_1 = 2$. Similarly, the absolute frequency of the value 6 is 2, or $AF_2 = 2$, the absolute frequency of 8 is 2, or $AF_3 = 2$, the absolute frequency of 9 is 1, or $AF_4 = 1$, etc. These values are shown in the 3rd column of the table following step 5.

STEP 2: The cumulative frequency for the data value 5 is 2or $CF_1 = 2$ because there are 2 values of 5, the cumulative frequency for the data value 6 is 4, or $CF_2 = 4$ because there are 2 values of 5 and 2 values of 6, etc. The cumulative frequencies for all 8 distinct data values are shown in the 4th column.

STEP 3: The values $Y_i = CF_i / (n+1)$ for each of the 8 distinct data values are shown in the 5th column of

the table. For example, $Y_1 = 2 / 16 = 0.125$, $Y_2 = 4 / 16 = 0.25$.

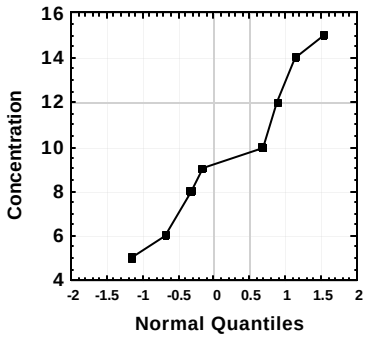
STEP 4: Note the standard normal distribution is symmetric about a mean of zero and furthermore that Table A.1 only gives quantiles for normal data values that are greater than or equal to zero. Hence, if the value of Y_i (from Step 3) is less than 0.5, we must compute $1 - Y_i$ and determine the quantile for that value. Note also the quantile for any value of Y_i less than 0.5 will be a negative number.

The normal (distribution) quantile of the distinct data value 5 is obtained from Table A.1as follows. From Step 3 and the following table we know that $Y_1 = 0.125$, which is less than 0.5. Hence, compute $1.00 - Y_1 = 1.00 - 0.125 = 0.875$. Then find 0.875 in the body of Table A.1 and read off the corresponding value from the left and top margins of the table. For 0.875 we find that value to be 1.15. Hence, the quantile for the data value 5 is $Q_1 = -1.15$. Similarly, for the distinct data value 6, we see from the table that $Y_2 = 0.25$, which is less than 0.5. Hence we compute $1 - Y_2 = 0.75$. We find 0.75 in the body of Table A.1 and read off the value 0.675 from the left and top margins (using linear interpolation). Hence, $Q_2 = -0.675$. The other values of Q_i are found similarly, except that when $Y_i \geq 0.50$, then we find the value of Y_i in the body of the table and read off the quantile (which is a positive value) from the left and top margins.

The values of $Q_1, Q_2, \ldots, Q_8$ are given in the 6th column of the following table.

STEP 5: Plot the $n' = 8$ pairs (x_i, Q_i). This plot follows. It appears the points are approximately linear, but it is not very conclusive as there are so few distinct data points; only 8 are present. Unless there are 20 or more distinct data points, the probability plots or the formal statistical tests in Section 2.6 are not decisive tools for deciding which distribution best fits the data.

i	**Individual x_i**	**Absolute Frequency AF_i**	**Cumulative Frequency CF_i**	**Y_i**	**Normal Quantiles Q_i**
1	5	2	2	0.1250	-1.15
2	6	2	4	0.2500	-0.675
3	8	2	6	0.3750	-0.319
4	9	1	7	0.4375	-0.157
5	10	5	12	0.7500	0.675
6	12	1	13	0.8125	0.387
7	14	1	14	0.8750	1.150
8	15	1	15	0.9375	1.534



If the data set for a metal of interest appears to be normally distributed for both the Navy site and background data sets, then consider using either the two-sample t test (Section 3.8) or the Satterthwaite two-sample t test (Section 3.9) to evaluate if the metal is a COPC. For all other

situations, we recommend that one of the nonparametric tests for the COPC be used (Sections 3.4, 3.5, 3.6, 3.7, and 3.10).

2.5.5 Interpreting Probability Plots

Several reasons exist why one may want to know whether a data set fits a single, hypothesized distribution:

1. The two-sample t test and the Satterthwaite two-sample t test discussed in Sections 3.8 and 3.9, respectively, require that site and background data sets be normally distributed.
2. Being able to demonstrate that a combined data set (see Section 2.2) fits a single distribution provides justification for using the combined data set in a test for COPC. A probability plot that suggests outliers may be present or that is not well fit by a straight line gives rise to speculation the data set may in fact be a combination of two quite different data sets. For example, one portion of the data set may be from a contaminated part of the site and the other portion from an uncontaminated part of the site. See Sections 2.5.6 and 3.1.1 for further discussion of this point.
3. Certain environmental processes can give rise to common theoretical distributions. If no other information is available about the historical processes that took place at the site, the distribution of data may give hints into historical events that may have occurred. For example, some natural processes tend to produce lognormal distributions (for example, size of pebbles, annual amounts of rainfall), while anthropogenic and site operations may give rise to mixtures of distributions (such as spills that move through soil layers by the action of rainfall over the years). If you can establish that a data set fits well a standard theoretical distribution, you can try to narrow the types of processes that might have generated the data to those that are compatible with the hypothesized distribution. This information may be helpful in selecting a valid background data set and in determining whether data sets (site or background) can be combined.

Probability plots can also be useful for identifying potential outliers. A data value (or a few data values) much larger or much smaller than the rest will cause the other data values to be compressed into the middle of the graph. If the plots do not exhibit a linear pattern, their nonlinearity will indicate the way in which the data do not fit the hypothetical distribution. This information is in addition to the statistical tests that distributions in Section 2.6 do not provide. Three typical distribution characteristics that will cause probability plots to deviate from a straight line are asymmetry (skewness), outliers, and heavy tails of the distribution. Helsel and Hirsch (1992, pages 30-33) describe these three conditions in detail.

2.5.6 Using Probability Plots to Identify Background

Another use of probability plots that has been proposed is to use a change in the slope or existence of an inflection point in the plotted line to indicate a break in the data set. The inflection point is said to identify a background threshold value that represents the upper

range of ambient conditions. In this section, we discuss some of the potential pitfalls of this use of probability plots.

As way of background, the probability plotting method for establishing background typically includes the following elements:

- Gaps or inflection points in the probability plot suggest multiple populations, including possible outliers. A straight-line plot with no gaps or inflection points indicates a single population. Probability plots should be used in conjunction with descriptive statistics (Section 2.3) for identifying ambient conditions.
- For the purpose of identifying the COPC for risk assessment, ambient (local background) conditions are defined as the range of concentrations associated with the population nearest the origin of the probability plot. This definition may be performed by inspection.
- Ambient data sets may be suspected of containing high measurements due to site activities, if the range of detected values is more than 2 orders of magnitude or the coefficient of variation (see Box 2.1) is greater than 1.

The following four subsections point out four potential problems in using probability plots to extract background data from a data set that may represent an entire installation-wide database. Also see Section 3.1.1 for further discussion.

1) Selection of the Hypothesized Distribution and the Interpretation of Hinge Points in the Probability Plot

Selection of the Hypothesized Distribution and the Interpretation of Hinge Points in the Probability Plot

Hinge points or inflection points in probability plots do not always indicate multiple populations. They may only indicate that the data do not follow the hypothesized distribution used in constructing the probability plot. To illustrate, the plot on the left-hand side of Figure 2.5-8 shows a probability plot of data from a lognormal distribution. However, the plot was constructed assuming the data had a normal distribution. Just looking at this plot we might conclude that it contains more than one population because of the breaks in the plotted points and the apparent distinct sections with different slopes. However, this conclusion is incorrect because the hinges and different slopes are only an indication that the data do not follow the hypothesized normal distribution used to construct the probability plot. To see this, look at the probability plot on the right-hand side of Figure 2.5-8. This plot was constructed using the same data, but assuming (correctly) that the data have a lognormal distribution. We see that the data now plot as a straight line.

This example illustrates that it is a good idea to construct probability plots for more than one underlying (assumed) distribution, e.g., both the normal and lognormal distributions. Also, breaks and hinge points in a probability plot should be a trigger to look more carefully at the data to determine if those features really indicate separate populations or if they can be explained in terms of site operational history, geological features or elements, location and time of sample collection, problems in the analytical laboratory, etc.

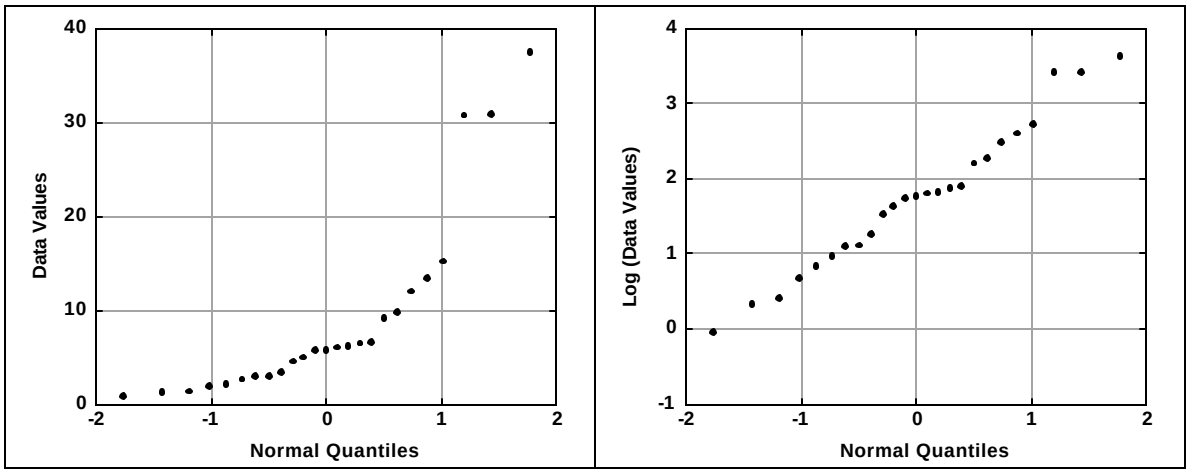


Fig. 2.5-8. Normal and Log-Normal Probability Plots of Log-Normal Data

2) Overlapping Populations

The lack of a hinge point does not necessarily mean only one population is present. This is illustrated in Figure 2.5-9 and Figure 2.5-10. Figure 2.5-9 shows boxplots of log-transformed aluminum concentrations for six different soil series (that is, six different populations). Figure 2.5-10 shows a normal probability plot of the pooled data from these six populations. Note that Figure 2.5-10 does not contain any hinge points, even though populations 3 and 4 (Figure 2.5-9) have very different distribution shapes and median values. Hinge points do not appear in Figure 2.5-10 because the six populations overlap, making them indistinguishable in the probability plot. Clearly, for overlapping populations such as illustrated in Figure 2.5-9, probability plots will not alert the user to the presence of the different populations.

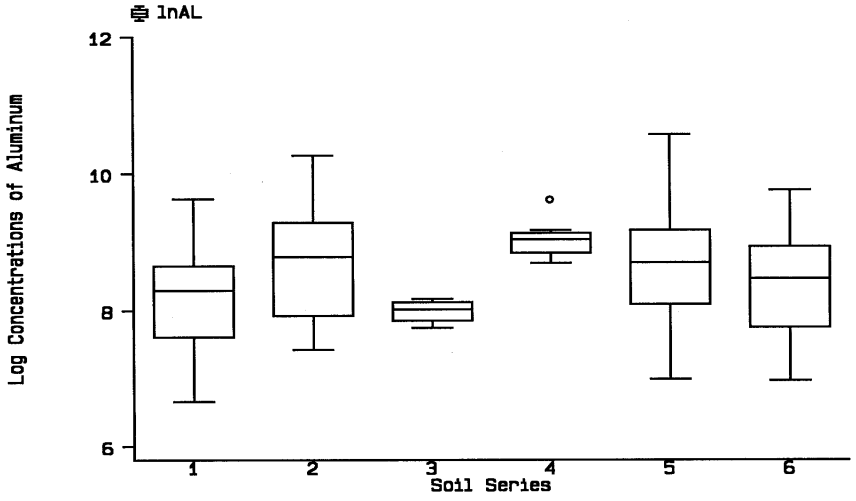


Fig. 2-5-9. Boxplots of Log-transformed Aluminum Concentrations in Six Different Soil Series

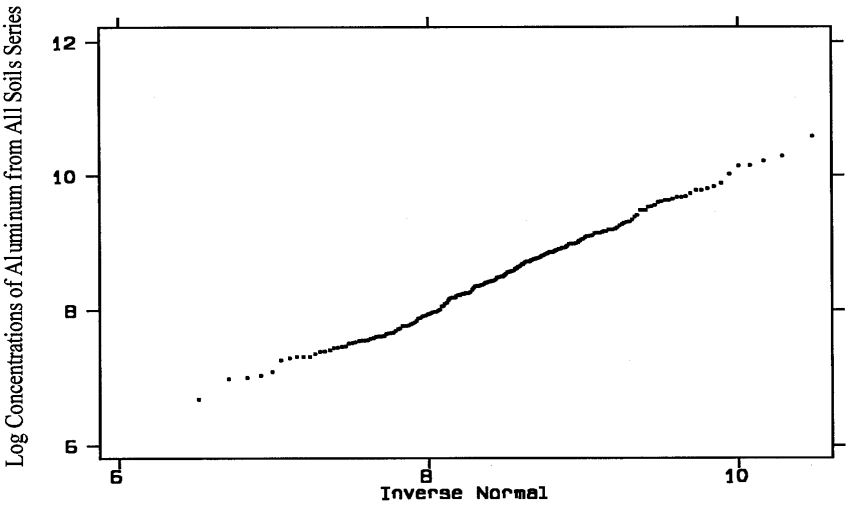


Fig. 2.5-10. Normal Probability Plot of Log-Transformed Aluminum Data from All Soil Series

But what if the only information available is a single data set and there is no way to know if it represents one or many populations, overlapping or not? Can probability plots provide useful information to determine the possibility of multiple populations? A good procedure to address these questions is to first use probability plots to screen a data set. If inflection points or hinges exist, then explore the differentiated data sets individually to see if process knowledge or physical evidence supports the hypothesis of distinct populations. If no hinges or infection points are obvious in the probability plot, explore further using supplemental information about site history or other data characteristics (such as soil type and location of data) that might suggest distinct populations are present. Use boxplots to visually explore the characteristics of the separate hypothetical data sets to see if the hypothesis of distinct populations can be supported by the data.

3) Hinge Points that Cut Off the Upper Portion of the Background Population

An important concern when using hinge points in probability plots to identify the background population is the technique can cut off the upper portion of the background population. To illustrate, we refer back to the soil data sets 1 and 2 of aluminum concentrations in Figure 2.5-9. Figure 2.5-11 is a normal probability plot of the data set formed by pooling data sets 1 and 2. Suppose that data set 1 is the ambient aluminum background distribution of interest and data set 2 is composed of aluminum data collected from a site being evaluated for possible contamination above ambient background. The probability plot in Figure 2.5-11 might be judged to contain a hinge point in the vicinity of the value 8.0. But looking at Figure 2.5-9 we see that 8.0 underestimates the upper range of the background data distribution. Therefore, use of the hinge point to define the background threshold would discount over half of the background data set. Consequently, the background threshold would be biased low. Hence, if each site data value were compared to this low background threshold value, the metal would be falsely identified as a COPC.

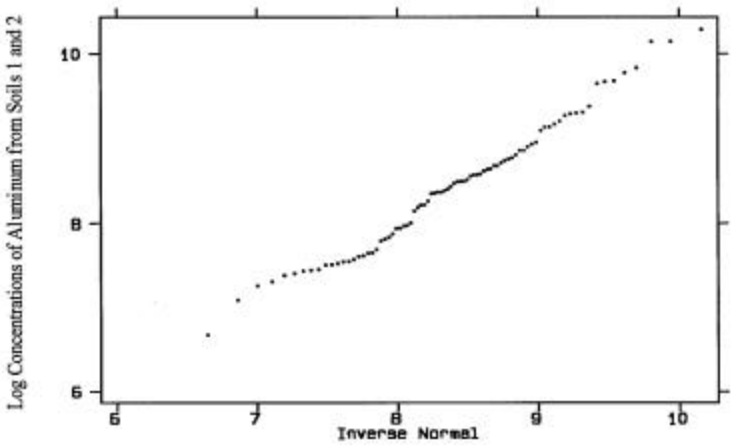


Figure 2.5-11. Normal Probability Plot of Log-Transformed Aluminum Data from the Combined Soil Series 1 and 2

4) Problems with Multiple Non-detects in Background and Site Data Sets

Two possible ways to handle non-detects before constructing a probability plot are:

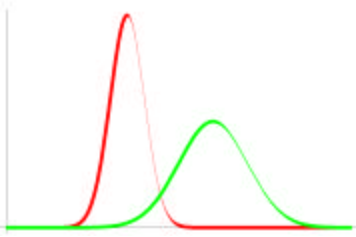
- replacing non-detects by one-half the sample quantitation limit (SQL) for each non-detect, or
- assigning all non-detects a dummy value at or below the lowest detected value

Rather than using one of these methods, it may be preferable instead to construct probability plots using the methods described in Akritas et al. (1994, page 227) or Michael and Schucany (1986, page 476, equation 11.8). These authors use state-of-the-art statistical procedures for properly constructing probability plots when multiple non-detects are present. However, these methods are somewhat complicated and their use for constructing probability plots for identifying COPC has not been evaluated.

Summary

In summary, the probability plotting approach outlined at the beginning of Section 2.5.6 is a simple way of graphically describing data but the interpretation of the plots is difficult. Thus, other graphical plots such as box plots should also be used to aid in interpreting the probability plots. Also, the interpretation of hinge points on these graphical plots is subjective. Different reviewers will disagree on whether a hinge point really exists or is just an artifact of the methodology.

2.6 Determining the Probability Distribution of a Data Set



Information about the location and shape of data distributions helps us analyze the data.

Why do we need to know which probability distribution best describes the data set?

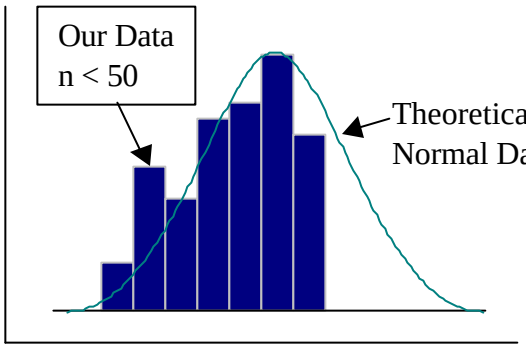
- To select the best statistical test for determining COPC
- To select the best statistical test for outliers
- To determine if the data should be transformed before applying tests for COPC or outliers
- To better model underlying processes that generated the data values

The selection of the best statistical test for testing whether a chemical is a COPC depends in part on whether the data set is normally distributed. For example, the two-sample t test requires the data to be normally distributed, but the Wilcoxon Rank Sum test does not (see Chapter 3.0 for these tests). In addition, we have discussed in Section 2.2 the shape and location of data sets should be similar before pooling the data sets. Also, most of the tests for outliers in Section 2.3 require the data set be normally distributed. If the data are not normally distributed, then the Walsh test for outliers may be used. Knowing the shape of the data set may also help to understand the environmental processes that had an impact on the data values. For example, if the data set appears to fit a normal distribution, this suggests that the concentrations are rather similar over the entire site (assuming representative samples were obtained), and that may help determine the origin and deposition process of the contamination.

In this section, we provide several methods for testing if the data are normally distributed. These tests can also be used to test whether the data set appears to fit a lognormal distribution. The procedure is to transform each datum to natural logarithms before conducting the outlier test. If the test indicates the transformed data are not normally distributed, the original (untransformed) data are not lognormally distributed.

2.6.1 Shapiro-Wilk W Test

The Shapiro-Wilk W test is highly recommended for testing whether the data have a normal distribution. It may also be used to test for a lognormal distribution, if the data are first transformed by computing the natural logarithm of each datum.



Can we say our data has a normal distribution?

Reason for Using the W Test

The W test is recommended in several EPA guidance documents (EPA 1992a and EPA 1996) and in many statistical texts (Gilbert 1987; Conover 1980). It is available in many software packages including GRITS/STAT (EPA 1992b) and DataQUEST (EPA 1997). The W test has been shown to have more power than other tests to detect when data are not from a normal or lognormal distribution. The W test should be conducted in conjunction with constructing normal and lognormal probability plots (Section 2.5.4) in order to more thoroughly evaluate whether the normal or lognormal distribution is an acceptable fit to the data.

Assumptions and Their Verification

An assumption that should be verified before the W test is used is that data values are independent and representative of the underlying population of possible measurements. This assumption is most likely to be valid if a suitable random sampling or systematic square or triangular grid sampling design is used to determine the sampling locations and if the sampling locations are not clumped or too close to each other. The procedure that was used to determine the sampling locations should be checked to verify that these conditions are fulfilled. If a suitable sampling design was not used, the data may not be representative of the underlying (site or background) population, in which case the W test results will not be meaningful.

Advantages and Disadvantages

The W test:

- requires the use of a special table of coefficients (Table A.6) and critical values (Table A.7)
- can only be conducted if the number of samples is less than or equal to 50 because the Table A.7 of critical values does not extend beyond n = 50
- is somewhat tedious to compute by hand, but it is easily conducted using the DataQUEST software
- should not be used if the data set contains non-detects

- may not have sufficient power to detect non-normality if the underlying distribution is only slightly different than the normal distribution or if the number of data in the data set is too small.

Table 2.1 shows the power of the W test to detect a lognormal distribution in the data, rather than a normal distribution. This table was obtained using computer simulations for which 1000 data sets of n measurements each were generated from lognormal distributions with various degrees of skewness (long tail towards high concentrations). Values of the power are provided in Table 2.1 for various numbers of samples (from 10 to 100) and lognormal distribution shapes, as indicated by the coefficient of variation (CV, which is the standard deviation divided by the mean). The CV range from 0.1 to 1.3. Lognormal distributions that are only slightly asymmetric will have a small CV, whereas highly skewed (asymmetrical) lognormal distributions have large CV.

The results in Table 2.1 show the W test does not have a high probability of differentiating a lognormal distribution from a normal distribution when the natural variability of the population is low (that is, when the CV is small that indicates an almost symmetrical distribution shape) and the number of data values, n, is small. For example, when n and CV are both small, say n = 20 and CV = 0.20, the probability that the W test will correctly reject the null hypothesis the background population is normally distributed is only 0.12; about one chance in 10. However, CV greater than 0.50 and sample sizes greater than 20 are typically encountered in establishing background conditions and determining COPC. The power of the W test for this range of values is 0.50 or greater and may be considered adequate. But it is clear from Table 2.1 the W test should not be relied on to detect non-normality, if fewer than 20 representative measurements have been obtained (unless the CV of the underlying distribution is substantially greater than 0.50).

Table 2.1. Power of the W Test to Reject the Null Hypothesis of a Normal Distribution when Underlying Distribution is Lognormal.

Power of W Test for Simulated Test Conditions													
	CV												
n	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	**1.1**	**1.2**	**1.3**
10	0.059	0.077	0.109	0.179	0.225	0.273	0.342	0.403	0.452	0.487	0.508	0.537	0.565
15	0.080	0.096	0.177	0.242	0.376	0.462	0.534	0.626	0.677	0.701	0.746	0.777	0.811
20	0.054	0.117	0.232	0.346	0.496	0.599	0.684	0.746	0.825	0.851	0.893	0.925	0.923
25	0.081	0.185	0.299	0.434	0.562	0.741	0.817	0.86	0.887	0.930	0.961	0.964	0.970
30	0.066	0.206	0.371	0.513	0.698	0.791	0.876	0.891	0.959	0.973	0.978	0.986	0.992
35	0.077	0.192	0.348	0.603	0.746	0.831	0.903	0.967	0.970	0.983	0.987	0.993	0.998
40	0.101	0.219	0.459	0668	0.826	0.903	0.957	0.972	0.992	0.996	0.998	0.997	0.996
60	0.135	0.349	0.608	0.832	0.946	0.972	0.995	0.997	0.999	1	0.999	1	1
70	0.112	0.363	0.706	0.883	0.961	0.989	0.999	1	1	1	1	1	1
80	0.127	0.396	0.732	0.921	0.987	0.997	0.999	0.999	1	1	1	1	1
90	0.171	0.448	0.79	0.941	0.992	0.999	1	1	1	1	1	1	1
100	0.156	0.551	0.811	0.970	0.993	0.999	1	1	1	1	1	1	1

The computations needed to conduct the W test and an example are provided in Box 2.19.

Box 2.19. Shapiro-Wilk W Test

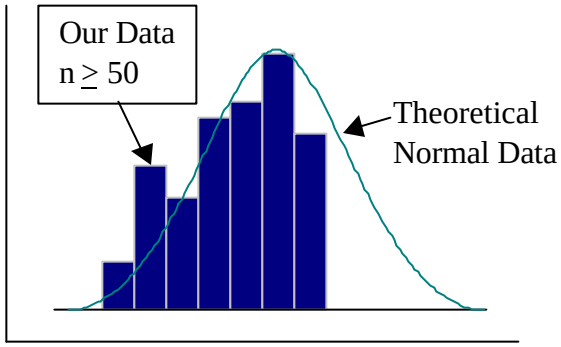
- Select the significance level, α, desired for the test, where $0 < \alpha < 0.5$. That is, select the probability, α, that can be tolerated of the W test declaring that the measurements in the data set are not from a normal distribution when in fact they are from a normal distribution.
- Compute the arithmetic mean of the n data: $\bar{x} = (x_1 + x_2 + \ldots + x_n) / n$
- Compute the denominator d of the W test statistic using the n data and $\bar{x}$:
  $d = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2$
- Order the n data from smallest to largest. Denote these "sample order statistics" by
  $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$.
- Compute k, where $k = n/2$ if n is an even integer and $k = (n-1)/2$ if n is an odd integer
- Turn to Table A.6 to obtain the coefficients $a_1, a_2, \ldots, a_k$ for the value of n.
- Compute the W test statistic
  $W = \{ a_1(x_{(n)} - x_{(1)}) + a_2(x_{(n-1)} - x_{(2)}) + \ldots + a_k(x_{(n-k+1)} - x_{(k)}) \}^2 / d$
- Conclude that the data set is not normally distributed if the value of W is less than the critical value given in Table A.7 for the selected significance level α.

Example:

- Suppose we select $\alpha = 0.05$
- Suppose there are n = 10 measurements in the data set:
  1.20, 0.13, 1.69, 1.05, 1.12, 0.45, 2.06, 0.60, 0.76, 1.37.
- The arithmetic mean of these data is
  $\bar{x} = (1.2 + 0.13 + 1.69 + 1.05 + 1.12 + 0.45 + 2.06 + 0.60 + 0.76 + 1.37) / 10$
  $= 1.04$
- The denominator d of the W test statistic using the n data and $\bar{x}$ is:
  $d = (1.2 - 1.04)^2 + (0.13 - 1.04)^2 + \ldots + (1.37 - 1.04)^2$
  $= 3.05$
- Order the n = 10 measurements from smallest to largest to obtain:
  0.13, 0.45, 0.60, 0.76, 1.05, 1.12, 1.20, 1.37, 1.69, 2.06
- Compute $k = n/2 = 10/2 = 5$ because n is an even integer.
- In Table A.6 we find that the k = 5 coefficients are
  $a_1 = 0.5739$, $a_2 = 0.3291$, $a_3 = 0.2141$, $a_4 = 0.1224$, $a_5 = 0.0399$
- Hence, the computed W statistic is:
  $W = \{0.5739(2.06 - 0.13) + 0.3291(1.69 - 0.45) + 0.2141(1.37 - 0.60)$
  $+ 0.1224(1.20 - 0.76) + 0.0399(1.12 - 1.05)\}^2 / 3.05$
  $= 0.989$

The critical value from Table A.7 for n = 10 and $\alpha = 0.05$ is 0.842. Hence, as 0.989 is not less than 0.842, we conclude the measurements appear to be normally distributed. The data do not provide convincing evidence the distribution of the measurements is not normal.

2.6.2 D'Agostino Test



Can we say our data has a normal distribution?

D'Agostino Test (D'Agostino 1971) may be used to test if the measurements are from a normal distribution.

Reason for Using the D'Agostino Test

The Shapiro-Wilk W test, discussed in Section 2.6.1, cannot be used if n > 50. However, D'Agostino's Test can be used when n ≥ 50. D'Agostino (1971) showed the performance of his test compares favorably with other tests.

Assumptions and Their Verification

The same comments provided in Section 2.6.1, regarding assumptions and their verification for applying the W test, also apply to the D'Agostino test.

Advantages and Disadvantages

The D'Agostino test:

- cannot be conducted if n < 50 or n > 1000
- requires the use of a special table of critical values to conduct the test (Table A.8)
- is tedious to compute by hand
- cannot be conducted if the data set contains non-detects
- may not have large power to detect non-normality if the underlying distribution is only slightly different than the normal distribution or if the number of data in the data set is small

The computations needed to conduct the test are provided in Box 2.20 along with an example.

Box 2.20. D'Agostino Test

- Select the significance level, α, desired for the test, where $0 < \alpha < 0.5$.
- Compute $s = \{[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2] / n\}^{1/2}$
- Order the n data from smallest to largest. Denote these sample order statistics by $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$.
- Compute $D = \{[1 - 0.5(n+1)]x_{(1)} + [2 - 0.5(n+1)]x_{(2)} + \ldots + [n - 0.5(n+1)]x_{(n)}\} / n^2 s$
- Compute $Y = (D - 0.282094) / (0.02998598 / n^{1/2})$
- Conclude the data are not from a normal distribution, if Y is less than the critical value $Y_{\alpha/2}$ or greater than the critical value $Y_{1-\alpha/2}$, that are found in Table A.8 for each value of n.

Example (from Gilbert 1987, page 161):

- Suppose we select $\alpha = 0.05$
- Suppose n = 115 and the computed value of s is $\{[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2] / 115\}^{1/2} = 0.4978$
- Then the value of $n^2 s$, the denominator of D, is $(115)^2(0.4978) = 6583$
- As $0.5(n+1) = 0.5(116) = 58$, and using the sample order statistics $x_{[i]}$, the numerator of D equals $\{[1-58]x_{(1)} + [2-58]x_{(2)} + \ldots + [115 - 58]x_{(115)}\} = 1833.3$
- Hence, $D = 1833.3 / 6583 = 0.2785$
- Hence, $Y = (0.2785 - 0.282094) / (0.02998798 / 115^{1/2}) = -1.29$
- From Table A.8, we find using linear interpolation that $Y_{0.025} = -2.522$ and $Y_{0.975} = 1.339$.
- Since –1.29 is not less than –2.522 and not larger than 1.339, we cannot conclude that the measurements are not normally distributed.

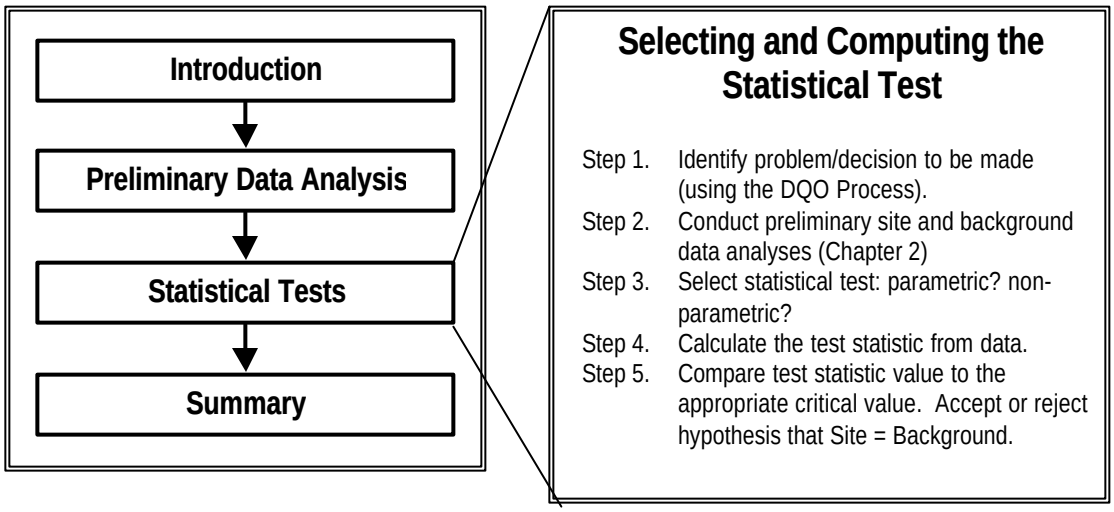
2.6.3 Other Tests

A large number of statistical tests, in addition to those discussed in Sections 2.6.1 and 2.6.2, could be used to test hypotheses about which probability distribution best fits a data set. These tests are commonly called *goodness-of-fit tests.* A thorough summary of the scientific literature on this topic, with many examples provided, is in D'Agostino and Stephens (1986). This book is suitable for someone who has some training in statistics. EPA (1986, Section 4.2) provides more easily understood descriptions of several tests, most of which can be conducted using the DataQUEST software (EPA 1997).

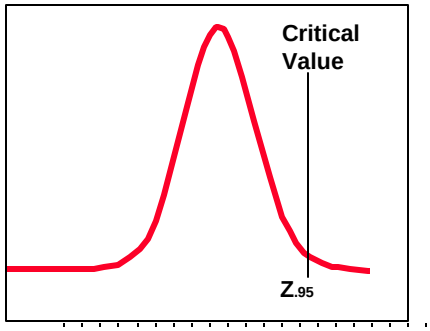
EPA (1996) recommends the use of the W test if the number of samples is less than 50. They recommend either the Filliben statistic or the studentized range test otherwise. The Filliben test (Filliben 1975) is not illustrated here because it is closely related to the W test and is a bit difficult to compute by hand, although it is easily computed using DataQUEST software, EPA (1997). EPA (1996, p. 4.2-2) highly recommends the studentized range test except when the data appear to be lognormally distributed. The test, illustrated in EPA (1996, p. 4.2-5), is simpler to compute than the W test and critical values needed for the test are available for sample sizes (n) up to 1000.

If several goodness-of-fit tests are applied to the same data set, the decisions indicated by the test (of whether the data fit the specified distribution) may differ among the tests. If so, that should be taken as an indication the data do not contain sufficient information to decide the issue with assurance.

3.0 STATISTICAL TESTS TO COMPARE SITE AND BACKGROUND



3.1 Selecting a Statistical Test



Is Test Statistic > $Z_{.95}$?

As discussed in Navy (1998), all decisions regarding whether Navy site or operation area concentration levels tend to be larger than background concentrations should consider the results of statistical tests. An initial, tentative selection of the most appropriate statistical test(s) to perform should be made during the DQO planning process. This selection should be based on the number of samples required for the various tests to achieve the specified performance goals (DQO), the particular distribution (normal or lognormal) expected of the data to be collected, and information in published statistical papers that demonstrate the performance of the candidate tests for various data distributions and contamination scenarios. However, after the new data have been collected and the preliminary graphical and distribution data analyses have been conducted as discussed in Chapter 2, a final selection of the statistical test(s) can be made.

The assumptions and advantages and disadvantages of each of the tests discussed in this chapter are provided in Box 3.1 to aid the reader in selecting the most appropriate statistical test(s). In this regard, note that the optimal selection of a test depends in part on whether

- the entire distribution of the observed measurements from the site is simply shifted to higher values than the observed distribution of background measurements, or

- the true concentrations in relatively small areas at the site are elevated relative to the true background concentrations, in which case only a small portion of the distribution of site measurements would be expected to be shifted to higher concentrations than the distribution of background measurements.

For the case of a simple shift, the two-sample t test, the Satterthwaite t test, and the Wilcoxon Rank Sum (WRS) test are the preferred tests. However, the Slippage test, Quantile test, and the two-sample test for proportion are better suited to identify metals that have elevated concentrations in only small areas at the site. All of these tests are discussed in Table 3.1 and later sections of this handbook.

All tests require that site and background measurements be independent (not spatially or temporally correlated) and representative of the underlying site and background populations. This assumption requires (1) an appropriate probability-based sampling design strategy be used to determine the location of soil samples to be collected, and (2) the soil samples are far enough apart in space and time that spatial and temporal correlations among concentrations at different locations are not present. Also, to help guard against the tests having power that is too low to reliably detect a COPC, the number of samples (data values) in both the background and site data sets for all the statistical tests should be at least 10 and, hopefully, more than 20.

Minor differences are noted between this handbook and Figure 11 in Navy (1998). That figure shows a flowchart for deciding which statistical tests should be conducted. Comments relevant to these differences are as follows:

- The Slippage test is not mentioned in Figure 11, but it is included in this handbook (Section 3.4). The test is included here because it is very simple to conduct and it has intuitive appeal, using the maximum observed background datum as a background threshold. In Figure 11 the Quantile test is used in place of the Slippage test. The Quantile test has somewhat greater power than the Slippage test to detect when a metal is a COPC and it is only slightly more complex to conduct.
- The two-sample test of proportions (Section 3.10) is not included in Figure 11. This test is discussed in this handbook because it useful when many non-detects are present in the site or background data sets.
- Figure 11 indicates the two-sample t test (Section 3.8) or the Satterthwaite two-sample t test (Section 3.9) should be computed on the logarithms of the data, if statistical analyses (Sections 2.5 and 2.6) suggest the data are skewed to the right and lognormally distributed. However, if those tests are conducted on the log-transformed data, the medians, rather than the means of the site and background populations, are being compared. The results of tests on the medians of skewed data sets do not necessarily apply to the means of those data sets. Hence, caution is needed in interpreting the test results.
- The flowchart in Figure 11 indicates that, if the data are normally or lognormally distributed, neither the Wilcoxon Rank Sum test (Section 3.6) nor the Gehan test (Section 3.7) should be used. It is recommended that these tests *should* be used, regardless of whether the data have a normal, lognormal, or some other distribution, particularly if non-detects are present in the data sets. The guidance provided here is that the Wilcoxon Rank

Sum or the Gehan test should be used with the following exception. The tests should not be used if strong evidence exists that the site and background data sets are normally distributed and essentially no non-detects are present in either data set. In that case, either the two-sample t test or the Satterthwaite two-sample t test should be used.

- Figure 11 indicates the Quantile test should be used if another test (for example, the Wilcoxon Rank Sum test) was used on the same set of background and site data and did not declare the metal to be a COPC. Hence, in some cases two tests will be performed on the same data. In that situation, the significance level, α, for the two tests combined will be about double that of the α level specified for each test. The guidance provided here is that both the Wilcoxon Rank Sum test and the Quantile test should be routinely used. This approach was recommended in EPA (1994b). Hence, the overall α level for both tests together should be specified and $\alpha/2$ used as the significance level for each of the two tests.

Box 3.1. Assumptions and Advantages/Disadvantages of Statistical Tests to Detect When Site Concentrations Tend to be Larger than Background Concentrations

Test Statistic	Assumptions	Advantages/Disadvantages
Slippage Test	• Objective is to test for differences in the right tail of the site and background concentration distributions • More less-than values are allowed than for other tests considered here • At least one detected (quantified) background measurement is present and it is larger than the largest less-than value • No assumptions are required with regard to the shape of site and background data concentration distributions.	Advantages: • Very simple to conduct the test • No distribution assumptions are necessary • Many less-than values are permitted • Can be used in conjunction (in tandem) with tests that focus on the detecting differences in the mean or median. Disadvantages: • Must be certain that the largest background datum is not a mistake or error. • May need large number of measurements to have adequate power to detect differences in site and background concentrations
Quantile Test	• Objective is to test for differences in the right tail of the site and background concentration distributions. • Less-than values are not among the largest r data values in the pooled set of site and background data. (See Section 3.5 for the definition of r.) • No assumptions are required with regard to the shape of the site and background data concentration distributions	Advantages: • Relatively simple to conduct the test • No distribution assumptions are necessary • Can have more power to detect differences in the right tail of site and background distributions than tests like the WRS, Gehan or Two-Sample t tests that focus on the mean or median. • Can be used in conjunction (in

		tandem) with tests that focus on detecting differences in the mean or median Disadvantages: • May need large number of measurements to have adequate power to detect differences in site and background concentrations • Test may be inconclusive if less-than values are present among the r largest data values.
Wilcoxon Rank Sum Test	• Objective is to test for differences in the medians of the site and background populations • Only one detection limit (all less-than values have the same value), which is less than the smallest detected datum. • No more than 40% of both the site and background data sets are less-than values • No assumptions are required with regard to the shape of the site and background data concentration distributions	Advantages: • No distribution assumptions necessary • In general, the test has more power to detect shift in site median than the two-sample t tests when the site and background data distributions are asymmetric (skewed to the right, to high concentrations). • Can be used in conjunction (in tandem) with Slippage and Quantile tests so that differences in the right tails of the site and background distributions, as well as differences in medians, can be detected Disadvantages: • Relatively more complex to compute by hand • Too many less-than values prevent use of the test
Gehan Test	• Objective is to test for differences in the medians of the site and background populations • All less-than values do *not* have the same value (multiple detection limits exist). • The censoring mechanism that generated the less-than values is the same for the site and background populations • No assumptions are required with regard to the shape of the site and background data concentration distributions	Advantages: • Can be used when multiple less-than values (multiple detection limits) are present • Same Advantages as for the WRS test Disadvantages: • Relatively complicated to compute by hand. • The performance of the test is not known as well as that of the WRS test. • Must assume the same censoring mechanisms apply to the site and background data

Two-Sample Test of Proportions	• Test may be used when more than 50% of the site or background data sets are less-than values • It is desired or necessary (because of many less-than values) to test for differences between the site and background populations using the proportion of concentrations that exceed some specified value • No assumptions are required with regard to the shape of the site and background data concentration distributions	Advantages: • No distribution assumptions are necessary • Relatively simple test to perform • Can be used when many less-than values are present Disadvantages: • A test based on proportions may not be what is really needed, that is, may really need to test for a shift in medians
Two-Sample t Test	• Objective is to test for differences in the means of the site and background populations • Both site and background data are normally distributed • No less-than values are present • Site and background data are expected or known to have the same total variance	Advantages: • Most powerful test for detecting a shift in the site mean from the background mean, if the site and background data are normally distributed Disadvantages: • The test requires a statistical evaluation of the assumption of equal total variances for the site and background populations • In general, the power will be less than that of the WRS test, if the data are not normally distributed • Normal distribution assumption is often violated • The results of the test can be affected by outliers • Not well suited for data sets that contain less-than values
Satterthwaite Two-Sample t Test	• Objective is to test for differences in the means of the site and background populations • Both site and background data have a normal distribution • No less-than values are present • Site and background data are expected or known to have unequal variances	Advantages: • Test can be used when the site and background distributions have unequal variances Disadvantages: • The test is relatively complicated to compute by hand • See Disadvantages of the two-sample t test

3.1.1 The Threshold Comparison Method

A method that has been proposed for identifying COPC is the *threshold comparison method.* The threshold comparison method consists of comparing the highest concentration (measurement) detected at the site with a concentration that represents the upper range of ambient (local background) conditions. While this approach has some intuitive appeal, two important questions must be addressed before that method is used:

- Question 1: "How is the background threshold value determined?"
- Question 2: "What is the likelihood that using the threshold comparison method will result in falsely concluding that a metal is a COPC?"

The following 6-step procedure has been proposed by some regulatory agencies to address Question 1. (Sections in this handbook that provide additional information for each step are indicated in parentheses.)

1. If possible, expand the background data set (Section 2.2)
2. Use a statistical test, such as the Shapiro-Wilk W test, to test the background data set for normality and lognormality (Section 2.6)
3. Compute descriptive statistics for the background data set (Section 2.3)
4. Construct a normal or lognormal probability plot of the data (the threshold comparison method refers to these plots as cumulative probability plots) (Section 2.5)
5. Use the probability plot to identify possible outliers (Section 2.4), as well as the set of data points nearest the origin that represents ambient conditions
6. Select the background threshold value as the value that represents the upper range of ambient conditions. One suggestion for selecting the threshold value is if the number of background measurements, m, is small, the threshold value may be the mean or an upper confidence limit on the mean. If m is large, the threshold value may be an upper percentile, such as the 95^{th} percentile or even the 99^{th} percentile.

An important limitation of this 6-step procedure occurs in Step 5. In that step, assuming the probability plot indicates one or more straight-line segments, the inflection or break points in the probability plot are used to help identify the background threshold value. While inflection points may indeed indicate separate underlying populations, no assurance is given that separate populations do indeed exist. This potential problem is discussed and illustrated in Section 2.5.

A second important limitation of the 6-step comparison method is the likelihood (probability) the comparison method (comparing the maximum site measurement to the background threshold value) will falsely declare that the metal is a COPC *will increase and eventually go to 1* as the number of site measurements, n, becomes large. In other words, even when the site and background concentration distributions are identical and, hence, the metal is not a COPC, the comparison method has a high probability of declaring the metal *is* a COPC if n is sufficiently large. The specific probabilities of making false positive decision errors for different values of n are given in Section 3.3.

The idea of comparing site measurements to some threshold value is not unique. It can also be found in EPA (1994b) and MARSSIM (1997). EPA (1994b) does not provide detailed guidance on how the background threshold value should be determined. It simply states the threshold value might be based on a site-specific risk assessment or an upper confidence limit for an upper percentile of the background distribution. EPA (1994b, Section 4.4.3) also states the threshold value might be determined by negotiation between regulators and the site owner or operator. Also, it should be used in conjunction with the WRS and Quantile tests (discussed in Table 3.1).

MARSSIM (1997) discusses the threshold approach in some detail. That publication focuses on final status radiological surveys for demonstrating compliance with risk-based standards when the radionuclide may occur in the background area. MARSSIM assumes the threshold value (denoted by $DCGL_{EMC}$, the Derived Concentration Guideline Limit for the Elevated Measurement Comparison) will be developed using exposure pathway models and that elevated concentrations will occur in relatively small areas at the site. As discussed in MARSSIM (1997, page 8-9), if any site measurement exceeds the threshold ($DCGL_{EMC}$), this is a flag or trigger for further investigation. The investigation may involve taking further measurements to determine the area and level of concentrations or assessing the models and methods used to determine the threshold value. It may also include an assessment of the consistency of the results obtained with the site history and other pertinent information. It should also be noted that the EMC test is always used in conjunction with other statistical tests recommended in MARSSIM. Furthermore, additional investigation is needed, regardless of the outcome of the other statistical tests. The other tests in MARSSIM (the WRS and Sign tests) are used to determine whether or not the site as a whole meets the risk release criterion. The EMC is used to screen individual measurements.

Based on the previous discussion, it is recommended the threshold comparison method:

- only be used as a trigger for additional investigation of whether the metal is a COPC
- never be the only criterion applied to determine if a metal is a COPC
- always be used in conjunction with
  - graphical plots in addition to probability plots
  - geochemical characteristics of the elements and geologic conditions, including the use of geochemical scatter plots and simple linear regression methods to look for associations of metals and their adsorbents (Navy 1998, Section 3.1.7)
  - one or more of the statistical tests described in Sections 3.1 through Section 3.10 of this handbook.

The Slippage test, that is discussed in Section 3.4, is somewhat related to the threshold comparison method, but it does not have the problems of determining the background threshold value and elevated false positive decision error rates. The Slippage test consists of simply counting the number of site measurements that exceed the maximum background measurement. If the count is sufficiently large (it must always be larger than 1), the test declares that the site distribution has *slipped* to values greater than that of background. In the context of this handbook, such a result provides evidence that the metal may be a COPC. The

Slippage test, as well as the related Quantile test (Section 3.5), may be regarded as statistically valid methods that might replace, or certainly supplement, the threshold comparison method.

3.2 Hypotheses Under Test

All tests discussed in this chapter are testing the following null and alternative hypothesis (denoted by H_o and H_a, respectively):

H_o: Navy site or operation concentrations do not tend to be larger in value than background concentrations, that is, the chemical of interest is not a COPC

H_a: Navy site or operation concentrations tend to be larger in value than background concentrations, that is, the chemical is a COPC.

When testing site versus background populations, the H_o is always initially assumed to be true. The H_o is rejected in favor of the H_a only when the data are sufficiently supportive of that decision. The burden of proof is demonstrating *beyond a reasonable doubt* that H_a is more likely to be true than H_o. This approach is the one used in recent publications on testing for compliance with background concentrations (EPA 1994b, MARSSIM 1997). Note this philosophy is also the one used in the United States legal court system (the accused person is assumed to be innocent until the evidence is sufficient to indicate beyond a reasonable doubt the person is really guilty). Hence, in using the H_o and H_a as defined previously, we are assuming before sampling is conducted that the chemical is not a COPC. Furthermore, the site data (as applied to the statistical test) must be convincing beyond a reasonable doubt before we conclude that the chemical is indeed a COPC. This approach requires that careful attention be given (via the DQO planning process) to collecting a sufficient number of representative concentration measurements from the site and background. This procedure should be followed so that the statistical test will have a sufficiently high probability of declaring the chemical is a COPC when in fact that is the case. If too few samples are collected, the statistical test may not have a sufficiently high probability of rejecting H_o and declaring truthfully that the chemical is a COPC.

You may ask "Why not interchange the H_o and H_a in order to be more protective of human health and the environment?" Interchanging H_o and H_a would give

H_o: Navy site or operation concentrations tend to be larger in value than background concentrations, that is, the chemical *is* a COPC.

H_a: Navy site or operation concentrations do not tend to be larger in value than background concentrations, that is, the chemical of interest is *not* a COPC.

These latter hypotheses are not used because statistical tests would have limited ability to correctly reject H_o (and declare that the chemical is not a COPC) unless Navy site concentrations were actually *less* than background concentrations, which seems to be an unreasonable requirement. In other words, the tendency would be for statistical tests to falsely conclude the chemical is a COPC. We note in passing that if one wants to test that a site is in compliance with a risk-based, *fixed threshold (upper limit)* concentration value (rather than with the distribution of background concentrations), then there is no problem using the latter hypotheses, that is, with using **H_o**: Chemical is a COPC, versus **H_a**: Chemical is not a COPC. However, testing for compliance with a risk-based threshold value is not the topic of this handbook.

3.3 Statistical Testing Approaches *Not* Recommended

This section describes two methods for comparing data that are not recommended for testing whether or not a chemical is a COPC. The methods are not acceptable because, as shown in following paragraphs, the probability the tests will give the wrong answer is too large. See Section 3.1.1 for a related discussion.

3.3.1 Comparing the Maximum Site and Maximum Background Measurements

One approach to test whether a chemical is a COPC is to compare the maximum site measurement with the maximum background measurement, using the following decision rule:

> If the maximum site measurement exceeds the maximum background measurement, then declare the chemical is a COPC; otherwise declare the chemical is not a COPC.

As discussed in O'Brien and Gilbert (1997), the following two key issues make this methodology unsuitable:

- **Issue 1:** Suppose the site and background areas have the same concentration distribution and, hence, the chemical is not a COPC. If *unequal* numbers of samples are measured at the site and background, the data set (site or background) with the most measurements has the higher probability of containing the maximum measurement among the two data sets.

- **Issue 2:** If the site and background truly have the same concentration distribution and if an *equal* number of samples are measured for the chemical for both the site and background, the probability is 0.50 the maximum measurement occurs in the site data set and 50% that it occurs in the background data set. Thus, the chance is 50% that the chemical will be declared to be a COCP, when in fact the chemical is at background levels on the site.

With regard to Issue 1, if the site and background distributions are identical (have the same shape and location) and the site data set has n measurements and the background data set has m measurements, the probability is $P = n/(n+m)$ that the site data set will have the largest measurement (assuming the measurements are independent and were obtained using simple random sampling). If $n = m$, then $P = 0.50$, as noted above in Issue 2.

Suppose, for example, that $n = 20$ and $m = 10$. That is, if twice as many site measurements as background measurements are obtained, then $P = 20/30 = 2/3$. That is, the probability is 0.67 using the decision rule (of comparing the maximum site measurement with the maximum background measurement) will lead to *incorrectly* declaring that the chemical of interest is a COPC.

Clearly, this decision rule is not acceptable because its performance in declaring whether or not a chemical is a COPC depends so critically on whether the site or background area has the most measurements. However, a simple and defensible decision rule is known that *can* be used to compare site measurements with the maximum background measurement to determine COPC. The test (decision rule) is called the Slippage Test, discussed in Section 3.4.

3.3.2 Comparing the Maximum Site Measurement to a Background Threshold

Another decision rule that might be used to decide if a chemical at the site is a COPC is:

> If one or more site measurements exceed the 95th percentile of the background distribution, declare the chemical of interest to be a COPC.

Suppose the site and background distributions are identical and, thus, the chemical is not a COPC. Then, if the previous decision rule is used, it can be shown the probability that one or more of n site measurements will exceed the 95th percentile is equal to $1 - (0.95)^n$, where 0.95 is the probability that any randomly drawn (representative) single site measurement is less than the 95th percentile of the background distribution. The expression $1 - (0.95)^n$ takes on the values shown in Box 3.2 for various values of n.

Box 3.2. Probabilities that One or More of n Site Measurements Will Exceed the 95th Percentile of the Background Distribution if the Site and Background Distributions are Identical

n	$1 – (0.95)^n$
1	0.05
2	0.10
5	0.23
8	0.34
10	0.40
12	0.46
21	0.67
64	0.96

For example, if the background and site distributions are identical and if n = 21 site measurements of the chemical are obtained, the probability that one or more of the site measurements will exceed the 95th percentile of the background distribution is 0.67. In other words, the probability of obtaining a false positive result (declaring the chemical is a COCP when that is not really the case) is 0.67. If more extensive sampling is conducted at the site, for example, if n = 64, the probability of falsely concluding the chemical is a COPC is 0.96.

The danger of using this type of decision rule is clear: the probability of making a false positive decision error can be unacceptably large when many site measurements are compared to a background threshold value.

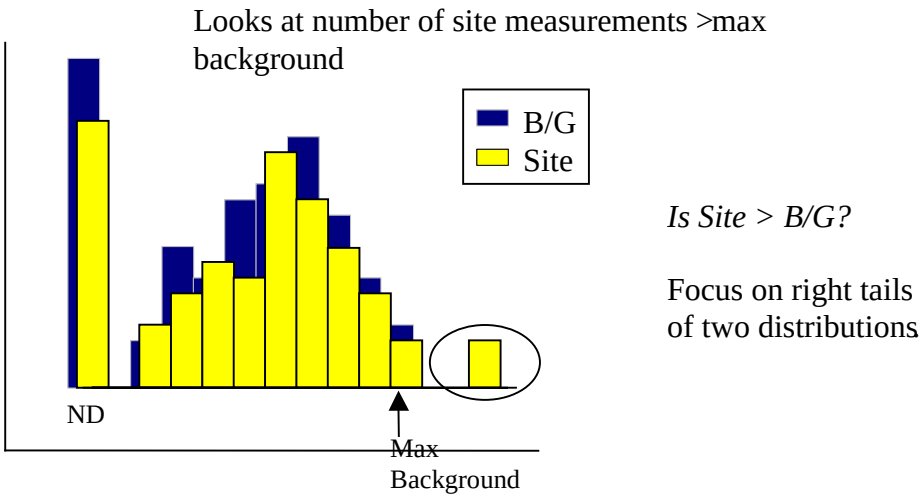
Threshold values, other than the 95th percentile, that might be used include the 90th or 99th percentiles. Also, the background mean, two times the background mean, or an upper confidence limit on the background mean might be suggested as appropriate threshold values. Regardless of which threshold value is selected, it will correspond to some percentile (perhaps unknown) of the background distribution. Hence, no matter which threshold value is used, the basic problem of too many false positive decision errors remains if site measurements are

individually compared to the threshold value. Only the specific probability of making a false positive decision error changes.

We note that the above decision rule is modified by using, say the 90^{th} percentile rather than the 95^{th} percentile, the probability that one or more of the n site measurements will exceed the 90^{th} percentile of background is $1 - (0.90)^n$ when the site and background distributions are identical. This example illustrates that the same formula, $1 - (\text{threshold percentile})^n$, is used for computing the probability of making a false positive decision error, regardless of what threshold percentile is selected.

3.4 Slippage Test

Site Contamination Scenario



Suppose certain factors give us reason to believe that operations at a Navy facility may have released small amounts of a contaminant to surface soil in a 1000 m^2 region (Region A) at the facility. Also, it is known that this particular contaminant is also present in soils in the natural environment in a defined background area located close to the Navy facility. The decision to be made is whether the concentration levels of this contaminant within Region A exceed the median for the natural background area. If so, the contaminant will be declared to be a COPC. Knowledge of site operations suggests that if releases of the contaminant did occur, the contamination may not be evenly spread across Region A, although most parts of the region are expected to have relatively low concentrations.

Role of the Data Quality Objectives Process

The DQO process was used to reach agreement with stakeholders and regulators regarding the methods that should be used to collect, handle, prepare and measure the soil samples. Consensus was reached that less-than measurements may frequently occur. It was also agreed that the decision of whether the chemical is a COPC should be made (at least in part) on the basis of only the larger site and background measurements. The slippage test was selected for this purpose because it uses only the largest few data values and does not require any assumptions about the underlying distribution of the site and background measurements. The assumptions that underlie the Slippage test are given in Box 3.1.

The stakeholders and regulators also decided to use the Wilcoxon Rank Sum (WRS) test in order to also look for differences in the medians of the site and background distributions as a criterion for deciding if the chemical is a COPC. The WRS test is described in Section 3.6

Advantages and Disadvantages

- The Slippage test consists of counting the number of site measurements that exceed the largest background datum and then comparing that count with a critical value from a special table (see Box 3.4). Hence, the slippage test is extremely easy to conduct.

- The Slippage test cannot be applied if the largest background datum is a less-than value. However, the test can be conducted in a straight-forward manner even if m-1 of the m background data are less-than values, as long as the largest background less-than value is less than the largest background detected value.

- As the Slippage test only uses the largest background measurement and the largest few of the site data measurements, it is important to verify these values are not mistakes or errors made during sample collection, handling, measurement or data handling. A test for outliers (Section 2.4) can be used to help decide if the largest values are unusually large, relative to what is expected based on an assumed distribution for the other measurements in the data set. If so, these outliers should be scrutinized to decide if they are mistakes or errors. To be safe it is a good idea to scrutinize suspiciously large values, even if the outlier test does not indicate they are outliers.

- If the number of samples is sufficiently large, a high probability exists that the Slippage test will detect when the right tail of the site distribution is shifted to higher concentrations than the right tail of the background concentration distribution.

- In general the Slippage test will not have high power to detect a shift in the *mean or median* of the site distribution relative to the mean or median of the background distribution. This situation occurs because the test looks at only the largest background measurement and the largest few site measurements.

- The Slippage test and the Quantile test (the latter test is discussed in Section 3.5) are closely related. However, the Slippage test is so simple to perform that it takes essentially no additional effort to conduct. It can be viewed as a quick test to see almost at a glance whether it is likely the chemical is a COPC. However, if the Slippage test fails to declare that a chemical is a COPC, this result should not be used to make a final conclusion that the chemical is not a COPC. Additional statistical testing, using the WRS test (Section 3.6) is needed.

- In general, the WRS test has better performance than the Slippage test to detect when the site concentrations are more or less uniformly greater across the entire site than background concentrations. The Slippage test performs better than the WRS test at detecting when only a portion of the site has concentrations much greater than the background area, assuming representative samples are collected from all regions of the site and background.

Guidance on Implementing the Slippage Test

The first step in implementing the Slippage test is to determine the number of site and background measurements, n and m, respectively, that are required for the test to have adequate power to declare (when it is true) the chemical of interest is a COPC. The required values of n and m depend not only on the required power, but also on the following design parameters:

1. The proportion, ε, of the site that has concentrations greater than background.

2. The amount that site concentrations tend to be larger than background concentrations.

3. The probability, α, that can be tolerated that the Slippage test declares the chemical is a COPC when in fact it *is not a COPC*.

4. The underlying distributions (for example, normal or lognormal) of the site and background concentration measurements.

Little information is present in the scientific literature concerning the best values of n and m for use in the Slippage test. However, Gilbert and Simpson (1990) provide enough information in their Table 1 and Figure 3 to provide the general guidance in Box 3.3. This box gives the approximate minimum number of measurements, n and m (for when n = m) that should be used in the Slippage test to achieve a power (probability) of approximately 0.80 and 0.90 for various values of ε. Their results are for the case where the tolerable value selected for α is between 0.025 and 0.05. Additional information on the power of the Slippage test is given in Figure 3 of Gilbert and Simpson (1990).

Box 3.3. Minimum Number of Samples (n and m) Required by the Slippage Test to Achieve a Power of Approximately 0.80 or 0.90 when a Proportion, e, of the Site has Concentrations Substantially Larger than Background (from Table 1 in Gilbert and Simpson 1990)

	Number of Required Measurements (n and m)	
e	**Power » 0.80**	**Power » 0.90**
0.10	60	75
0.15	40	50
0.20	30	35
0.25	25	30
0.30	15	25
0.35	15	20
0.40	15	20
0.45	10	15
0.50	10	10
0.60	10	10

It is important to note the following three points.

- The results in Box 3.3 are for the case where all site concentrations (in the ε region) are larger than *any* true background concentration. If it is suspected that some site concentrations in the ε region will be similar in value to background concentrations, but a few will be definitely larger than background measurements, the n and m in Box 3.3 will be too small to detect this small difference.

- If a value of α smaller than 0.025 is selected, the number of samples in Box 3.3 would have to be increased for the Slippage test to retain a power of 0.80 or 0.90. If a value of α larger than 0.05 is selected, the number of samples in Box 3.3 could be decreased somewhat and the Slippage test would still have a power of 0.80 or 0.90

- If site and background measurements have already been collected and the budget does not allow for additional samples, the information in Box 3.3 can be used to approximately determine if a power of 0.80 and 0.90 can be achieved with the available number of measurements. If not, the data by themselves may not contain enough information for the Slippage test to make a confident decision about whether the chemical is a COPC. Other sources of reliable information, such as expert knowledge about Navy operations at the site, should be used to the maximum extent in making COPC decisions.

Box 3.4 gives the procedure for conducting the Slippage test. Examples are provided in Boxes 3.5 and 3.6.

Box 3.4. Procedure for Conducting the Slippage Test

1. Specify the probability, α, that can be tolerated of the Slippage test incorrectly declaring the site concentrations tend to be larger than the background concentrations. The probability α can only be selected to be 0.01 or 0.05 because critical values for conducting the test are only available for those two values of α (Step 7 below). α is the probability the test will incorrectly declare the chemical is a COPC. NOTE: When both the Slippage test and the WRS test are conducted, the α level of the combined tests will be approximately the sum of the α level selected for each test.
2. Specify the values of ε and of the power $(1 - \beta)$ the stakeholders and regulators have decided are important for the Slippage test.
3. Determine the approximate minimum value of n = m from Box 3.3.
4. Collect the n = m samples and measure the chemical of interest in each sample. Some of the measurements may be less-than values.
5. Determine the value of the largest *detected* background measurement. In making this determination, ignore all less-than values that may be present in the background data set.
6. Count the number, K, of detected *site* measurements that are larger than the largest detected background measurement. In making this determination, ignore all less-than values in the site data set.
7. If α was selected as approximately 0.01, determine the critical value K_c from Table A.9. If α was selected as approximately 0.05, determine K_c from Table A.10. Note that the value of K_c depends on n and m.
8. If K is larger than the critical value K_c, declare the site concentrations for the chemical of interest tend to be larger than the background concentrations for that chemical, that is, the chemical is a COPC.

Box 3.5. Example 1 of the Slippage Test

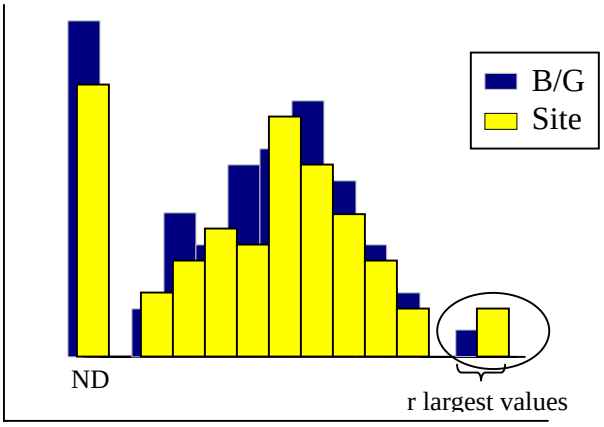
1.0 Suppose $\alpha = 0.01$ is selected
2.0 Suppose $\varepsilon = 0.50$ and a desired power of 0.80 are selected.
3.0 The approximate minimum number of measurements needed is n = m = 10 (from Box 3.3).
4.0 Suppose the following representative measurements of the chemical of interest are obtained (listed in order from smallest to largest):
Background Data: 23, 36, 37, 37, 44, 57, 60, 61, 61, 79
Navy Site Data: 15, 15, 20, 29, 30, 39, 60, 89, 90, 100
5.0 The value of the largest background measurement is 79.
6.0 K = 3 detected site measurements are larger than 79.
7.0 Entering Table A.9 with n = m = 10, we find the critical value K_c is 6
8.0 Hence, the Slippage test declares that evidence is insufficient to declare the chemical is a COPC because K = 3 is not larger than $K_c = 6$.
9.0 However, do *not* conclude that the chemical is *not* a COPC. Instead, also conduct the WRS test (Section 3.6) on these data.

Box 3.6. Example 2 of the Slippage Test

1.0 Suppose $\alpha = 0.05$ is selected
2.0 Suppose $\varepsilon = 0.30$ and a desired power of 0.80 are selected.
3.0 The approximate minimum number of measurements needed is n = m = 15 (from Box 3.3).
4.0 Suppose the following 30 representative measurements of the chemical of interest are obtained (listed in order from smallest to largest):
Background Data: <3, <3, <4, <7, <7, <8, 8, 15, <16, <16, <17, <17, 22, <24, <25
Navy Site Data: <5, <10, 11, 13, <22, 23, <24, <36, <40, 70, 89, <100, 115, 200, <300
5.0 The value of the largest detected background measurement is 22.
6.0 K = 5 detected site measurements are larger than 22.
7.0 Entering Table A.10 with n = m = 15 we find the critical value K_c is 4.
8.0 Hence, the Slippage test declares the chemical is a COPC because K = 5 is larger than $K_c = 4$.
9.0 Normally, the WRS test would also be performed to confirm the results of the Slippage test. However, the data sets contain so many less-than values the WRS test cannot be computed (see Section 3.6). The Gehan test (Section 3.7) should be used in place of the WRS test.

3.5 Quantile Test

Looks at number of site data among the largest data in pooled site and background data set.



Is Site > B/G?

Focus on right tails of two distributions.

Site Contamination Scenario

The site contamination scenario described for the Slippage test in Section 3.4 also applies to the Quantile test. A need exists to determine if the concentrations of the chemical of interest within, say, a 1000 m^2 region (Region A) at the facility tend to be greater than those in a defined background area. If so, the chemical should be declared a COPC. Knowledge of site operations suggests that if releases of the contaminant did occur, the contamination may not be evenly spread across Region A, although most parts of the region are expected to have relatively low concentrations. This situation suggests that the Quantile test is appropriate, although the WRS test should also be performed.

Role of the Data Quality Objectives Process

We assume the stakeholders and regulators used the data quality objectives (DQO) process to determine the methods that should be used to collect, handle, prepare and measure the soil samples. Consensus occurred that some less-than measurements would happen and the decision of whether the chemical of interest is a COPC should be made using the Quantile test in combination with the WRS test (Section 3.6).

The Quantile test was selected because (1) it is a valid test regardless of the underlying distribution of the site and background data, (2) the test looks for differences in the right tails of the site and background concentration distributions, and (3) the test complements the WRS test in the sense that the WRS test is good at detecting shifts in the medians and means.

Advantages and Disadvantages

- The Quantile test is a close cousin to the Slippage test. It consists of looking at the largest r measurements in the pooled site and background data sets and counting the number of those r measurements that are from the site. If k or more of the r measurements are site measurements, the Quantile test declares the chemical is a COPC. The Quantile test focuses on comparing the right tails of the site and background distributions rather than comparing the median or mean of the two distributions. For this reason, the Quantile test should always be used in tandem with the WRS test because the WRS test focuses on looking for differences in means and medians.

- Any number of less-than values are permitted in the site and background data sets, as long as all less-than values are smaller than the smallest of the r largest detected measurements in the pooled data set.

- In general, the WRS test has better performance than the Quantile test to detect when the site concentrations are more or less uniformly greater across the entire site than background concentrations. The Quantile test performs better than the WRS test in detecting when only a portion of the site has concentrations greater than the background area (assuming a sufficient number of representative samples are collected from all regions of the site and background)

- Use of the Quantile test does not require knowing the underlying concentration distribution of the chemical of interest. For example, the measurements need not be normally or lognormally distributed.
- Box 3.1 provides a summary of the advantages and disadvantages of the Quantile test.

- The procedure for conducting the Quantile test is shown in Box 3.9. Boxes 3.10 and 3.11 provide two examples of its use.

Guidance on Implementing the Quantile Test

As with other tests discussed in this handbook, the first step in implementing the Quantile test is to determine the number of site and background measurements, n and m, respectively, required for the test to have adequate power to declare (when it is true) the chemical of interest is a COPC. Also, in common with the Slippage test, the required values of n and m also depend on the

- proportion, ε, of the site that has concentrations greater than background
- amount that site concentrations tend to be larger than background concentrations
- probability, α, that can be tolerated of the Quantile test declaring on the basis of measurements, the chemical is a COPC when in fact it *is not* a COPC
- underlying distribution (for example, normal or lognormal) of the site and background concentration measurements.

EPA (1994b, Tables A.2, A.3, A.4, and A.5) provides information on the values of n and m required for the Quantile test to achieve prescribed power of the Quantile test to correctly declare a chemical is a COPC. A portion of those results are summarized here in Boxes 3.7 and 3.8. These boxes show the approximate number of site and background measurements needed (n = m) for the Quantile test to have a power (probability) of approximately 0.80 and 0.90 to correctly declare that a chemical is a COPC. These results are for the case where the tolerable probability, α, of the Quantile test incorrectly declaring, on the basis of measurements, the chemical is a COPC is specified by the stakeholders and regulators to be either 0.01, 0.025, 0.05, or 0.10. The results in Boxes 3.7 and 3.8 were obtained assuming the measurements are normally distributed. If it is suspected that measurements are skewed to the right and perhaps have a lognormal rather than a normal distribution, the number of samples should probably be increased somewhat to achieve the 0.80 and 0.90 power levels.

The number of measurements in Box 3.7 are those for which approximately 85% of the actual (true) site concentrations (in the ε portion of the site) are larger than the vast majority of background concentrations. The number of measurements in Box 3.8 are for the case where many site and background concentrations in the ε region will be similar in value, but about 5% of the site concentrations in the ε region are larger than the vast majority of background concentrations. The number of measurements are larger in Box 3.8 than in Box 3.7 because the results in Box 3.8 are for the case where site concentrations tend to be only *slightly* larger than background concentrations. Hence, it takes more information (measurements) to achieve the same power to detect differences.

The Quantile test can be computed using the software EnvironmentalStats for S-Plus (Millard 1997).

Box 3.7. Minimum Number of Measurements (n and m, n = m) Required by the Quantile Test to Achieve a Power of Approximately 0.80 or 0.90 When a Proportion, e, of the Site has Concentrations *Distinctly Larger* than Background Concentrations*

a :	**0.01**		**0.025**		**0.05**		**0.10**	
Power:	**0.80**	**0.90**	**0.80**	**0.90**	**0.80**	**0.90**	**0.80**	**0.90**
e = 0.10	>100	>100	100	>100	80	100	55	70
e = 0.20	55	60	40	40	35	40	25	35
e = 0.30	25	30	20	25	20	20	15	15
e = 0.40	20	25	15	20	15	15	10	15
e = 0.50	15	20	15	15	10	10	10	10
e = 0.60	10	15	10	10	10	10	10	10
e = 0.70	10	10	10	10	10	10	10	10
e = 0.80	10	10	10	10	10	10	10	10
e = 0.90	10	10	10	10	10	10	10	10
e = 1.0	10	10	10	10	10	10	10	10

***** n = m were obtained for the case where the normal site concentration distribution is shifted to the right of the normal background concentration distribution by the amount $\Delta/\sigma = 4$ (Tables A.2, A.3, A.4, and A.5 in EPA, 1994b). α is the probability (selected by stakeholders and regulators) that can be tolerated of the Quantile test incorrectly declaring, on the basis of the measurements, the chemical is a COPC.

Box 3.8. Minimum Number of Measurements (n and m, n = m) Required by the Quantile Test to Achieve a Power of Approximately 0.80 or 0.90 When a Proportion, e, of the Site has Concentrations *Somewhat Larger* than Background Concentrations*

a:	**0.01**		**0.025**		**0.05**		**0.10**	
Power;	**0.80**	**0.90**	**0.80**	**0.90**	**0.80**	**0.90**	**0.80**	**0.90**
e = 0.10	>100	>100	>100	>100	>100	>100	>100	>100
e = 0.20	>100	>100	>100	>100	>100	>100	>100	>100
e = 0.30	>100	>100	>100	>100	>100	>100	>100	>100
e = 0.40	>100	>100	>100	>100	>100	>100	>100	>100
e = 0.50	>100	>100	>100	>100	>100	>100	>100	>100
e = 0.60	>100	>100	>100	>100	>100	>100	>100	>100
e = 0.70	>100	>100	100	>100	75	>100	70	>100
e = 0.80	>100	>100	75	>100	60	>100	50	>100
e = 0.90	>100	>100	60	100	50	100	40	100
e = 1.0	>100	>100	50	75	50	75	30	75

* n = m were obtained for the case where the normal site concentration distribution is shifted to the right of the normal background concentration distribution by the amount $\Delta/\sigma = 1$ (Tables A.2, A.3, A.4, and A.5 in EPA, 1994b). α is the probability (selected by stakeholders and regulators) that can be tolerated of the Quantile test incorrectly declaring on the basis of the measurements that the chemical is a COPC.

Box 3.9. Procedure for Conducting the Quantile Test

1.0 Select the probability, α, that can be tolerated of the Quantile test incorrectly declaring the site concentrations tend to be larger than background concentrations. The probability α may be selected to be 0.01, 0.025, 0.05, or 0.10. NOTE: When both the Quantile test and the WRS test are conducted, the α level of the combined tests will be approximately the sum of the α levels selected for each test.

2.0 Specify the values of ε and of the power ($1 - \beta = 0.80$ or 0.90) desired for the test.

3.0 Use the specified values of ε and power in Box 3.7 to determine the approximate number of site and background measurements needed. Box 3.8 may be used if it is important to detect site concentrations that are only slightly larger than background.

4.0 Collect the n = m samples and measure the chemical of interest in each sample. Some of the measurements may be less-than values. If samples have already been collected and measured, verify their number is in agreement with Box 3.7 or Box 3.8. Collect additional samples, if necessary.

5.0 List from smallest to largest the pooled site and background measurements. The total number of pooled measurements is n + m.

6.0 Using the values of n and m, enter Table A.11, A.12, A.13, or A.14 (depending on whether α was selected to be 0.01, 0.025, 0.05, or 0.10, respectively) to find the values of r and k needed to conduct the Quantile test.

7.0 Determine from the ordered list of pooled site and background measurements if k or more of the largest *detected* r measurements are site measurements. (Note: ignore any less-than values when determining the largest detected r measurements). If so, the Quantile test indicates the chemical is a COPC. If not, the test indicates the measurements are insufficient for the Quantile test to conclude the chemical is a COPC. The WRS test should be computed.

Box 3.10. Example 1 of the Quantile Test

1.0 Suppose $\alpha = 0.05$ is selected.

2.0 Suppose $\varepsilon = 0.50$ is selected and a power of 0.80 is needed to detect when site concentrations are distinctly larger than background concentrations.

3.0 For these values of α, ε, and power, we see in Box 3.7 that a minimum of n = m = 10 measurements are required for the Quantile test.

4.0 Suppose the 20 measurements are as follows (the same data as was used to illustrate the Slippage test in Box 3.5):

Background Data: 23, 36, 37, 37, 44, 57, 60, 61, 61, 79

Navy Site Data: 15, 15, 20, 29, 30, 39, 60, 89, 90, 100

5.0 The 20 pooled and ordered background and site data are (S and B indicate Site and Background, respectively):

S S S B S S B B B S B B S B B B B S S S

15, 15, 20, 23, 29, 30, 36, 37, 37, 39, 44, 57, 60, 60, 61, 61, 79, 89, 90, 100

6.0 As $\alpha = 0.05$ was selected in Step 1, we find from Table A.13 for n = m = 10 that r = k = 4.

7.0 Among the largest r = 4 measurements in the pooled measurements (79, 89, 90, and 100), 3 are from the site. Hence, since 3 < k, that is, 3 < 4, the Quantile test indicates the measurements are insufficient to conclude the chemical is a COPC. The WRS test should be performed.

Box 3.11. Example 2 of the Quantile Test

1.0 Suppose $\alpha = 0.01$ is selected.

2.0 Suppose $\varepsilon = 0.50$ and a power of 0.80 is needed to detect when site concentrations are distinctly larger than background concentrations.

3.0 For these values of α, ε, and power, we see in Box 3.7 that n = m = 15 measurements are required for the Quantile test.

4.0 Suppose the data are as follows:

Background Data: <3, <3, <4, <7, <7, <8, 8, 15, <16, <16, <17, <17, 22, <24, <25

Site Data: <5, <10, 11, 13, <22, 23, <24, <36, <40, 70, 89, 100, 115, 200, 300

5.0 The 30 pooled and ordered background and site data are:

B B B S B B B B S S S B B B B

<3, <3, <4, <5, <7, <7, <8, 8, <10 11, 13, 15, <16, <16, <17,

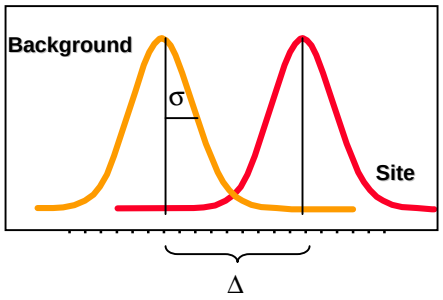
B S B S B S B S S S S S S S S

<17, <22, 22, 23, <24, <24, <25, <36, <40, 70, 89, 100, 115, 200, 300

6.0 As $\alpha = 0.01$ was selected in Step 1, we find from Table A.11 for n = m = 15 that r = k = 6.

7.0 Among the largest r = 6 detected measurements (70, 89, 100, 115, 200, 300), all 6 are from the site. Hence, since k (that is, 6) of the largest 6 (that is, r) measurements are from the site, the Quantile test indicates the chemical is a COPC.

3.6 Wilcoxon Rank Sum (WRS) Test

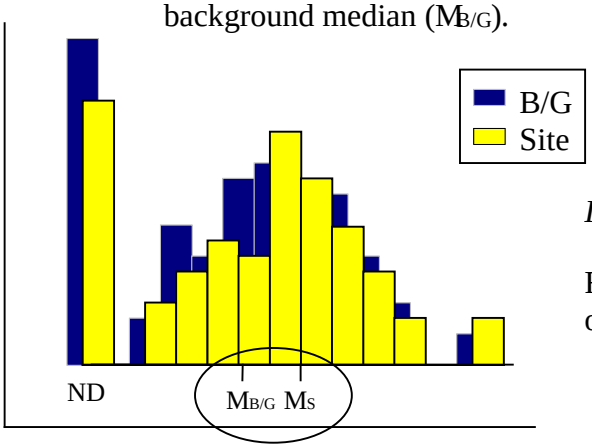


Is the site data distribution shifted to the right of the background data distribution by an important amount Δ?

Site Contamination Scenario

The site contamination scenario developed by the stakeholders and regulators during Steps 1 and 2 of the DQO process based on expert knowledge and all available past data was:

Asks if the site median (M_S) is larger than the background median ($M_{B/G}$).



Is site > background

Focus on medians of two distributions

If contamination from Navy site operations has occurred, it would probably be homogeneously distributed throughout the region within the site being investigated (Region A) rather than occurring in *hot spots* within that region.

Role of the Data Quality Objectives Process

Stakeholders and regulators also used the DQO planning process to agree

- on the methods that should be used to collect, handle, prepare, and measure the soil samples
- it was unlikely that more than 40% of the measurements would be less-than values
- that both the WRS test and the Quantile test should be conducted

- on the value of design parameters for determining the number of site and background measurements needed (see the section entitled "Guidance on Implementing the WRS Test" that follows).

The WRS test was selected because

- it is valid regardless of the underlying probability distribution of the site and of the background measurements

- the performance (power) of the test (in detecting when the median site concentration is shifted to the right of the median background concentration) is known from theory and practice to be as high or higher than other statistical tests that test for shifts in averages.

The Quantile test was selected to be conducted with the WRS test because it has more power (better performance) than the WRS test to detect when only a portion of Region A at the Navy site has concentrations greater than background. Hence, using the Quantile test in conjunction with the WRS test will improve the probability of detecting either uniform or non-uniform contamination greater than background.

WRS Test Assumptions and Their Verification

The underlying assumptions of the WRS test are:

- The measurements obtained from the soil samples from the Navy site and the background area are independent (not correlated). This assumption requires (1) that an appropriate probability-based sampling design strategy be used to determine the location of soil samples for collection, and (2) the soil samples are spaced far enough apart that a spatial correlation among concentrations at different locations is not present.

- The underlying probability distribution of the measurements in Region A has the same shape (variance) as the probability distribution for the background area. This assumption implies the two distributions are the same, except the distribution for Region A may be shifted to higher concentrations than the distribution for the background area. This assumption of equal variances should be evaluated using descriptive statistics and graphical plots of the Region A and background data (see Sections 2.3 and 2.5).

The assumptions that underlie the use of the WRS test are summarized in Box.3.1.

Advantages and Disadvantages

- If less-than values occur, all of them must have the same detection limit (the same less-than value), and that detection limit must be less than the smallest detected measurement. If multiple less-than values are scattered throughout the set of measurements, then the Gehan test (Section 3.7) should be used instead of the WRS test.

- The WRS test should not be used if more than 40% of the site or background data sets are less-than values. The measurement laboratories should be instructed to report actual measurements for all soil samples, whenever possible, even if the reported measurements are negative. Although negative concentrations cannot occur in nature, negative *measurements* can occur, due to measurement uncertainties, when true concentrations are very close to zero.

- The WRS test does not place large importance (weight) on the larger site and background measurements. It uses and considers *all* measurements, rather than focusing on the largest measurements as is done by the Slippage test and the Quantile test.

- The WRS test should be used in conjunction with the Quantile Test so that either uniform contamination or non-uniform contamination can be detected with greater probability.

- The software EnvironmentalStat for S-Plus (Millard 1997) can be used to compute the WRS test and the Quantile test.

Guidance on Implementing the WRS Test

To implement the WRS test, determine the number of site and background measurements to collect, denoted by n and m, respectively. A formula for computing n and m when the WRS test will be used is given in EPA (1994b, Equation 6.3, page 6.3). This sample-size formula requires inputs on

1. The acceptable probability, α, that the WRS test will *incorrectly* declare that the chemical is a COPC. Often, α is set at a value in the range of 0.01 to 0.10.

2. The required power (probability) the WRS test should have to *correctly* declare that the chemical is a COPC when that is in fact the case.

3. The amount Δ/σ (in units of standard deviation, σ) by which the site median concentration exceeds the background median concentration that must be detected with the required power.

4. The proportion of the total number of site and background soil samples that will be collected in the background area. If this proportion is 0.50, then n = m.

When n = m is desired (the usual case), a formula for determining the number of site and background measurements is given in MARSSIM (1997, Equation 5-1, page 5-28). However,

rather than use the formulas in EPA (1994b) or MARSSIM (1997), it is simpler to pick out n and m from Box 3.12 (which is Table 5.3 in MARSSIM 1997, page 5-30) if it is desired to have n = m. The values of n = m in Box 3.12 were obtained using Equation 5-1 in MARSSIM (1997) and then increasing that value by 20% to account for uncertainties and the likelihood that missing or unusable measurements will occur.

Box 3.12. Number of Site and Background Samples Needed to Use the Wilcoxon Rank Sum Test*

	a = 0.01					a = 0.025					a = 0.05					a = 0.10					a =0.25				
	Power					Power					Power					Power					Power				
D/s	0.99	0.975	0.95	0.90	0.75	0.99	0.975	0.95	0.90	0.75	0.99	0.975	0.95	0.90	0.75	0.99	0.975	0.95	0.90	0.75	0.99	0.025	0.05	0.10	0.25
0.1	5452	4627	3972	3278	2268	4827	3870	3273	2846	1748	3972	3273	2726	2157	1355	3278	2846	2157	1655	964	2268	1748	1355	964	459
0.2	1370	1163	998	824	570	1163	973	823	665	440	998	823	685	542	341	824	685	542	416	243	570	440	341	243	116
0.3	614	521	448	370	256	521	436	369	298	197	448	369	307	243	153	370	298	243	187	109	256	197	153	109	52
0.4	350	297	255	211	148	297	248	210	170	112	255	210	175	139	87	211	170	139	106	62	146	112	87	62	30
0.5	227	193	166	137	95	193	162	137	111	73	166	137	114	90	57	137	111	90	69	41	95	73	57	41	20
0.6	161	137	117	97	67	137	114	97	76	52	117	97	81	64	40	97	78	64	19	29	67	52	40	29	14
0.7	121	103	88	73	51	103	86	73	59	39	88	73	61	48	30	73	59	48	37	22	51	39	30	22	11
0.8	96	81	69	57	40	81	68	57	46	31	69	57	48	38	24	57	46	38	29	17	40	31	24	17	8
0.9	77	66	58	47	32	65	55	46	38	25	56	48	39	31	20	47	38	31	24	14	32	25	20	14	7
1.0	64	55	47	39	27	55	46	39	32	21	47	39	32	26	16	39	32	25	20	12	27	21	16	12	6
1.1	55	47	40	33	23	47	39	33	27	18	40	33	28	22	14	33	27	22	17	10	23	18	14	10	5
1.2	48	41	35	29	20	41	34	29	24	16	35	29	24	19	12	29	24	19	15	9	20	16	12	9	4
1.3	43	36	31	26	18	36	30	26	21	14	31	26	22	17	11	26	21	17	13	8	18	14	11	8	4
1.4	38	32	28	23	16	32	27	23	19	13	28	23	19	15	10	23	19	15	12	7	16	13	10	7	4
1.5	35	30	25	21	15	30	25	21	17	11	25	21	18	14	9	21	17	14	11	7	15	11	9	7	3
1.6	32	27	23	19	14	27	23	19	16	11	23	19	16	13	8	19	16	13	10	6	14	11	8	6	3
1.7	30	25	22	18	13	25	21	18	15	10	22	18	15	12	8	18	15	12	9	6	13	10	8	6	3
1.8	28	24	20	17	12	24	20	17	14	9	20	17	14	11	7	17	14	11	9	5	12	9	7	5	3
1.9	26	22	19	15	11	22	19	16	13	9	19	16	13	11	7	16	13	11	8	5	11	9	7	5	3
2.0	25	21	18	15	11	21	18	15	12	8	18	15	13	10	7	15	12	10	8	5	11	8	7	5	3
2.25	22	19	16	14	10	19	16	14	11	8	16	14	11	9	6	14	11	9	7	4	10	8	6	4	2
2.5	21	18	15	13	9	18	15	13	10	7	15	13	11	9	6	13	10	9	7	4	9	7	6	4	2
2.75	20	17	15	12	9	17	14	12	10	7	15	12	10	8	5	12	10	8	6	4	9	7	5	4	2
3.0	19	16	14	12	8	16	14	12	10	6	14	12	10	8	5	12	10	8	6	4	8	6	5	4	2
3.5	18	16	13	11	8	16	13	11	9	6	13	11	9	8	5	11	9	8	6	4	8	6	5	4	2
4.0	18	15	13	11	8	15	13	11	9	6	13	11	9	7	5	11	9	7	6	4	8	6	5	4	2

*Power is the probability the WRS test correctly declares that the chemical is a COPC.

Box 3.13 describes the steps to perform the WRS test when $n < 20$ and $m < 20$. Box 3.14 provides an example. Box 3.15 describes how to conduct the WRS test when $n \geq 20$ and $m \geq 20$, and Box 3.16 provides an example of that procedure.

Implementation Hints

The use of a triangular grid sampling design is suitable if the starting point of the grid is determined at random, and if the grid nodes (where samples are collected) are spaced far enough apart for the measurements to be independent. It is also necessary for the grid pattern not to coincide with a pattern of contamination in soil in such a way such that the estimated average concentration determined from the measurements is biased high or low. The use of a simple random sampling (SRS) design, where all locations are equally likely to be chosen, would also be an acceptable design. However, SRS can also result in relatively large portions of the area that have no locations to be sampled.

Box 3.13. Procedure for Conducting the Wilcoxon Rank Sum (WRS) Test when the Number of Site and Background Measurements is Small ($n < 20$ and $m < 20$)

1. Specify the probability, α, that can be tolerated of the WRS test incorrectly declaring that the site concentrations tend to be larger than the background concentrations, that is, of the test incorrectly declaring the chemical is a COPC. NOTE: When both the WRS and Quantile test are conducted, the α level of the combined tests will be approximately the sum of the α levels selected for each test.
2. Specify the value of Δ/σ and of power, where Δ/σ is the magnitude of the difference in median site and background concentrations that must be detected by the WRS test with the specified power. The notation Δ/σ indicates the shift is expressed in units of standard deviation (σ) of the underlying background and site concentration distributions for the chemical of interest.
3. Use the specified values of α, Δ/σ, and power in Box 3.12 to determine the number of site and background measurements needed when it is desired to have n equal to m. If having equal n and m is not desired, use Equation 6.3 in EPA (1994b) and increase that value by 20% to guard against missing or unusable measurements.
4. Collect the n and m samples and measure them for the chemical of interest, some of which may be less-than values. If measurements are available from past sampling efforts, verify their number is at least as large as the number indicated in Box 3.12. Collect additional samples, if necessary, to achieve the required number of samples.
5. List and rank the pooled set of n + m site and background measurements from smallest to largest, keeping track of which measurements came from the site and which came from the background area. Assign the rank of 1 to the smallest value among the pooled data, the rank of 2 to the second smallest value among the pooled data, and so forth.

   If a few measurements are tied (identical in value) assign the average of the ranks that would otherwise be assigned to the tied observations. If several measurement values have ties, average the ranks separately for each of those measurement values.

   If a few less-than values occur (say, <10%), and if all such values are less than the smallest detected measurement in the pooled data set, handle the less-than values as tied at an arbitrary value less than the smallest detected measurement. Assign the average of the ranks that would otherwise be assigned to these tied less-than values (the same procedure as for tied detected measurements).

   If between 10% and 40% of the pooled data set are less-than values, and all are less than the smallest detected

measurement, use the WRS test procedure is Box 3.15, even if n and m are less than 20. NOTE: The procedure in Box 3.15 is for the case where m and n are both of size 20 or larger. That procedure will provide only an approximate test if it is used when n and m are both smaller than 20. In that case, decisions of whether the chemical is a COPC should not be made until additional information is obtained by taking more samples and using a more sensitive measurement method.

6. Calculate the sum of the ranks of the *site* measurements. Denote this sum by R, then calculate W as follows:

$$W = R - n(n+1) / 2$$

7. Use the values of n and m and α to enter Table A.15 to find the critical value w_α , where α has been specified in Step 3 above. Table A.15 can only be used if α has been chosen to be 0.05 or 0.10.

   If $W > nm - w_\alpha$ the WRS test indicates the site concentration distribution is shifted to the right of the background concentration distribution, that is, that the chemical is a COPC.

8. If the WRS test declares the chemical is *not* a COPC, this conclusion may indicate (1) the chemical is indeed not a COPC, or (2) the assumptions that underlie the WRS test are not valid for the site and background measurements, or (3) an insufficient number of measurements (n and m) were obtained for the WRS test to detect the difference that actually exists in site and background concentration distributions.

   An evaluation should be made of the possibility the causes in items 2 or 3 may have resulted in the WRS test declaring the chemical is not a COPC. Review the DQO planning process records to make sure the number of samples (n and m) collected agree with what was determined at that time to be necessary to detect a possible difference between site and background measurements that was considered important. For case 3, the shift in the concentration distribution may in fact be smaller than the shift selected by the stakeholders, in which case no additional measurements are needed.

   Also, update the estimated number of site and background measurements needed by calculating the variance of the previous measurements and use that variance in the equation in Step 4 of Box 3.3-7 in EPA (1998). Collect additional samples if needed.

Box 3.14. Example of the Wilcoxon Rank Sum (WRS) Test when the Number of Site and Background Measurements is Small (n < 20 and m < 20)

Suppose, using the site contamination scenario given above, a need is present to determine if a chemical in surface soil in Region A on the Navy site is a COPC.

1. Suppose α was specified to be 0.05.
2. Suppose Δ / σ and the power were specified to be 1.5 and 0.95, respectively. That is, the stakeholders and regulators specified that if the median of the site concentration distribution is shifted by the amount Δ / σ greater than the median background distribution, then enough measurements should be obtained so that the WRS test has a power of 0.95 of detecting that fact.
3. Using these values of α, Δ / σ, and power to enter Box 3.12, we find that n = m = 18 measurements are needed for the WRS test.
4. Then, 18 samples from both the site and the background area were collected using a suitable probability-based sampling design (for example, simple random sampling or sampling at the nodes of a square or triangular grid) and measurements made of the chemical of interest on each sample. Suppose the measurements were:

   Background Data: 22, 32, 9, 12, 3, 7, 11, 2, 9, 11, 13, 16, 20, 25, <1, <1, 17, 21
   Site Data : 24, 33, 5, 9, 36, <1, 10, 50, 9, 19, 15, 10, 28, 9, 3, 15, 4, 19

5. Next, the data are pooled together and listed from smallest to largest. The ranks of the site data are determined (the site and background data and ranks are denoted by S and B, respectively):

	B	B	S	B	S	B	S	S	B	S	S	S	B	B	S	S	B	B
Data:	<1	<1	<1	2	3	3	4	5	7	9	9	9	9	9	10	10	11	11
Rank:	2	2	2	4	5.5	5.5	7	8	9	12	12	12	12	12	15.5	15.5	17.5	17.5

	B	B	S	S	B	B	S	S	B	B	B	S	B	S	B	S	S	S
Data:	12	13	15	15	16	17	19	19	20	21	22	24	25	28	32	33	36	50
Rank:	19	20	21.5	21.5	23	24	25.5	25.5	27	28	29	30	31	32	33	34	35	36

6. Sum the ranks of the site measurements to obtain R = 2 + 5.5 + 7 + … + 34 + 35 + 36 = 350.5. Hence,

   $$W = R - n(n+1) / 2 = 350.5 - 18(19) / 2 = 179.5$$

7. Enter Table A.15 with $\alpha = 0.05$ and n = m = 18 to obtain $w_{0.05} = 110$.

   We compute $nm - w_\alpha = 18 \times 18 - 110 = 214$. Therefore, $W < nm - w_\alpha$, that is, 179.5 < 214. The WRS has indicated the evidence in the data is insufficient to declare the chemical is a COPC.

As the WRS did not declare that the chemical is a COPC, the DQO process notes are reviewed to make sure the number of measurements specified to meet the α, Δ/σ, and power requirements were indeed obtained. Also, to update the estimated number of site and background measurements needed, calculate the variance of the previous 36 measurements and use that variance in the equation for n and m (n = m) in Step 4 of Box 3.3-7 in EPA (1998). If the number of samples computed using that equation exceeds the number used in the WRS test, collect the indicated number of new site and background samples.

Box 3.15. Procedure for Conducting the Wilcoxon Rank Sum (WRS) Test when the Number of Site and Background Measurements is Large ($n \geq 20$ and $m \geq 20$)

1. Specify the probability, α, that can be tolerated of the WRS test incorrectly declaring that the site concentrations tend to be larger than the background concentrations, that is, of the test incorrectly declaring the chemical is a COPC. NOTE: When both the WRS test and Quantile test are conducted, the α level of the combined tests will be approximately the sum of the α levels selected for each test.
2. Specify the value of Δ/σ and of power, where Δ/σ is the magnitude of the difference in average site and background concentrations that must be detected by the WRS test with the specified power. The notation Δ/σ indicates the shift is expressed in units of standard deviation (σ) of the underlying background and site concentration distributions for the chemical of interest.
3. Use the specified values of α, Δ/σ, and power in Box 3.12 to determine the number of site and background measurements needed when it is desired to have n equal m. If no need is present to have equal n and m, use Equation 6.3 in EPA (1994) and increase that value by 20% to guard against missing or unusable measurements.
4. Collect the n and m samples and measure them for the chemical of interest, some of which may be less-than values. If measurements are available from past sampling efforts, verify their number is at least as large as the number indicated in Box 3.12. Collect additional samples, if necessary to achieve the required number of samples.
5. List and rank the pooled set of n + m site and background measurements from smallest to largest, keeping track of which measurements came from the site and which came from the background area. Assign the rank of 1 to the smallest value among the pooled data, the rank of 2 to the second smallest value among the pooled data, and so forth.

   If < 40% of the measurements in the pooled data set are tied (identical in value) assign the average of the ranks that would otherwise be assigned to the tied observations. If several measurement values exist for which ties occur, average the ranks separately for each of those measurement values.

   If < 40% of the pooled data set are less-than values and if all such values are less than the smallest detected measurement in the pooled data set, handle those less-than values as being tied at an arbitrary value less than the smallest detected measurement. Assign the average of the ranks that would otherwise be assigned to this group of tied values (the same procedure as for detected measurements that are tied). NOTE: The total number of tied detected measurements and tied less-than values should not exceed 40% of the total number of measurements.

   If more than 40% of the pooled data are less-than values, then do not use the WRS test. The Gehan test should be used instead (Section 3.7).

6. Calculate the sum of the ranks of the site measurements. Denote this sum by R.
7. Calculate

$$w_{1-\alpha} = n(n+1)/4 + z_{1-\alpha}\,[n\,(n+1)\,(2n+1)/24]^{1/2}$$

   where $z_{1-\alpha}$ is the $100(1-\alpha)$ percentile of the standard normal distribution, which is tabulated in Table A.1. For example, if $\alpha = 0.05$, then $z_{1-\alpha} = z_{0.95} = 1.645$ from Table A.1.
8. The WRS test declares that the chemical is a COPC if $R > w_{1-\alpha}$.

Box 3.16 Example of the Wilcoxon Rank Sum (WRS) Test when the Number of Site and Background Measurements is Large ($n \geq 20$ and $m \geq 20$)

1. Suppose α is specified to be 0.01.
2. Suppose Δ / σ and the power were specified to be 1.8 and 0.95, respectively. That is, the stakeholders and regulators specified that if the median of the site concentration distribution is $\Delta / \sigma = 1.8$ (in units of standard deviation, σ) units greater than the median background distribution, enough measurements should be obtained so the WRS test has a power of 0.95 of detecting that fact.
3. Using these values of α, Δ / σ, and power to enter Box 3.12, we find that n = m = 20 measurements are needed for the WRS test, where n and m are the number of site and background measurements, respectively.
4. Then 20 samples from both the site and the background areas were collected using a suitable probability-based sampling strategy, for example, simple random sampling. Suppose the measurements were (listed in increasing magnitude):

   Background Data: <10, <10, <10, <10, 12, 15, 15, 18, 22, 26, 27, 29, 29, 29, 55, 60, 77, 90, 101, 150
   Site Data: <10, <10, <10, 25, 27, 27, 36, 36, 99, 101, 103, 140, 145, 150, 180, 190, 199, 200, 250, 300

5. Next, the data are pooled together and listed from smallest to largest. Then the ranks of the site data are determined (the site and background data and ranks are denoted by S and B, respectively).

	B	B	B	B	S	S	S	B	B	B	B	B	S	B	B	S	S	B	B	B	S
Data:	<10	<10	<10	<10	<10	<10	<10	12	15	15	18	22	25	26	27	27	27	29	29	29	36
Rank:	4	4	4	4	4	4	4	8	9.5	9.5	11	12	13	14	16	16	16	19	19	19	21

	S	B	B	B	B	S	B	S	S	S	S	B	S	S	S	S	S	S	S	
Data:	36	55	60	77	90	99	101	101	103	140	145	150	150	180	190	199	200	250	300	
Rank:	22	23	24	25	26	27	28.5	28.5	30	31	32	33	34	35	36	37	38	39	40	

6. The sum of the ranks of the site data is $R = 4 + 4 + 4 + 13 + 16 + \ldots + 39 + 40 = 507.5$.
7. Also,

$$w_{0.99} = n(n+1)/4 + z_{0.99}[n(n+1)(2n+1)/24]^{1/2}$$
$$= 20(21)/4 + 2.33[20(21)(41)/24]^{1/2}$$
$$= 167.4$$

   where $z_{0.99} = 2.33$ is the 99th percentile of the standard normal distribution, that is found in Table A.1.
8. As $R > w_{0.99}$, that is, $507.5 > 167.4$, the WRS test declares the chemical is a COPC.

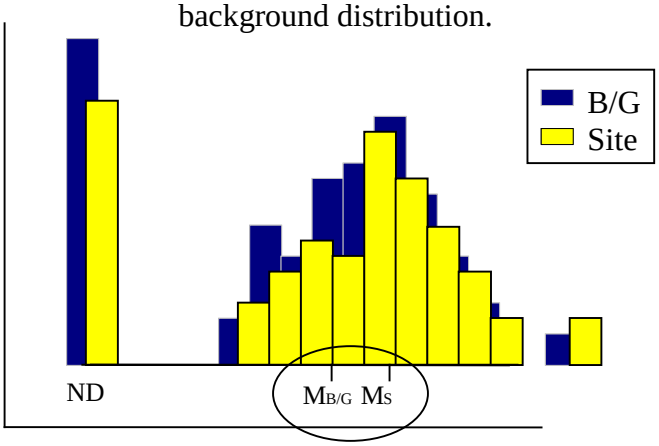
3.7 Gehan Test

Site Contamination Scenario

The site contamination scenario is the same as for the WRS test in Section 3.6.

Role of the Data Quality Objectives Process

Asks if the site distribution is shifted to the right of the background distribution.



Is site > background?

Focus on medians of two distributions

Stakeholders and regulators used the DQO planning process to agree that:

- the site and background data sets are likely to contain multiple less-than values of possibly different magnitudes, that is, all less-than values will not have the same detection limit.

- both the Gehan and Quantile tests should be used.

The Gehan test was selected instead of the WRS test because less-than values with different detection limits were expected to occur. The assumptions that underlie the use of the Gehan test are given in Box 3.1.

Advantages and Disadvantages

- The Gehan test can be used when the background or site data sets contain multiple less-than values with different detection limits.

- The Gehan test is somewhat tedious to compute by hand.

- If the censoring mechanisms are different for the site and background data sets, then the test results may be an indication of this difference in censoring mechanisms rather than an indication that the chemical is a COPC.

Implementation and Guidance

- The Gehan test is one of several nonparametric tests that have been proposed to test for differences between two sites when the data sets have multiple censoring points. Among these tests, Palachek et al. (1993) indicate they selected the Gehan test primarily because it was the easiest to explain, because the several methods generally behave comparably, and because the Gehan test reduces to the WRS test, a relatively well-known test to

environmental professionals. Palachek et al (1993) used their computer code to conduct Gehan tests on data from the Rocky Flats Environmental Technology Site near Denver, CO. They recommend using the Gehan test rather than a more complicated procedure involving replacement of non-detects by a value such as one-half the detection limit, testing for distribution shape and variance, and then conducting appropriate t tests or the Wilcoxon Ranks Sum test.

- The number of samples (measurements) needed from the site and from background to conduct the Gehan test may be approximated using the method described for the WRS test in Section 3.6. The procedure for conducting the Gehan test, when $n \geq 10$ and $m \geq 10$, is given in Box 3.17. An example of the test is given in Box 3.18. If $n < 10$ or $m < 10$, the procedure in Box 3.19 may be used to conduct the Gehan test.

Box 3.17. Gehan Test Procedure when m and n are Greater than or Equal to 10.

1. Specify the probability, α, that can be tolerated of the Gehan test incorrectly declaring that the site median is larger than the background median, that is, of the test incorrectly declaring that the chemical is a COPC.
2. Specify the value of Δ/σ and the power, where Δ/σ is the magnitude of the difference in median site and background concentrations that must be detected by the Gehan test with the specified power. The notation Δ/σ indicates the shift is expressed in units of standard deviation (σ) of the underlying background and site concentration distributions for the chemical of interest. Recall that an underlying assumption is that the variances of the site and background data for the chemical are the same.
3. Use the specified values of α, Δ/σ, and the power in Box 3.12 to determine the number of site and background measurements needed when it is desired to have n equal m. If no need exists to have equal n and m, use Equation 6.3 in EPA (1994a) and increase that value by 20% to guard against missing or unusable measurements.
4. Collect the n and m samples and measure them for the chemical of interest, some of which are expected to be less-than values. If measurements are available from past sampling efforts, verify that their number is at least as large as the number indicated in Box 3.12. Collect additional samples if necessary to achieve the required number of samples.
5. List the combined m background and n site measurements, including the less-than values, from smallest to largest, where the total number of combined samples is N = m + n. The less-than symbol (<) is ignored when listing the N data from smallest to largest.
6. Determine the N ranks, $R_1, R_2, \ldots, R_N$, for the N ordered data values using the method described in the example given below (Box 3.18).
7. Compute the N scores, $a(R_1), a(R_2),\ldots,a(R_N)$ using the formula $a(R_i) = 2R_i - N-1$, where i is successively set equal to 1, 2, …, N.
8. Compute the Gehan statistic, G, as follows:

$$G = \frac{\sum_{i=1}^{N} h_i \, a(R_i)}{\left\{ mn \sum_{i=1}^{N} [a(R_i)]^2 / [N(N-1)] \right\}^{1/2}} \qquad (1)$$

for which h_i = 1 if the i^{th} datum is from the site population
= 0 if the i^{th} datum is from the background population
N = n + m
$a(R_i)$ = $2R_i - N - 1$, as indicated above.

9. The Gehan test declares the chemical is a COPC if $G \geq Z_{1-\alpha}$, where $Z_{1-\alpha}$ is the $100(1-\alpha)^{th}$ percentile of the standard normal distribution, which is obtained from Table A.1. Otherwise, the test declares the evidence is not strong enough to conclude that the chemical is a COPC.

Box 3.18. Example of the Gehan Test

1. Suppose $\forall$ was specified to be 0.05.
2. Suppose Δ / σ and the power were specified to be 2.0 and 0.90, respectively. That is, the stakeholders and regulators specified that if the median of the site concentration distribution is greater than the median background distribution by the amount $\Delta / \sigma = 2.0$ (in units of standard deviation, σ), enough measurements should be obtained so the Gehan test has a power of 0.90 of detecting that fact.
3. Using the specified values of Δ / σ and power in Box 3.12, we find that n = m = 10 measurements are needed to conduct the Gehan test.
4. The 10 samples from the site and the background area were collected using a suitable probability-based sampling design (for example, simple random sampling or sampling at the nodes of a square or triangular grid) and measurements were made of the chemical of interest on each sample. Suppose the measurements are:

Background: 1 <4 5 7 <12 15 18 <21 <25 27
Site: : 2 <4 8 17 20 25 34 <35 40 43

5, 6 and 7. Use the following procedure to determine the N = 20 ranks $R_1, R_2, \ldots, R_{20}$ and the 20 scores $a(R_i)$. Refer to Table 1 below as you go through the steps.

Table 1. Calculations to Determine the Ranks, R_i, and the Scores, $a(R_i)$

Data	h_i	$Index_i$	d_i	e_i	R_i	$a(R_i)$	Data	h_i	$Index_i$	d_i	e_i	R_i	$a(R_i)$
1	0	0	1	0	4	-13	18	0	0	8	3	12.5	4
2	1	0	2	0	5	-11	20	1	0	9	3	13.5	6
< 4	1	1	2	1	4.5	-12	<21	0	1	9	4	8	-5
< 4	0	1	2	2	4.5	-12	<25	0	1	9	5	8	-5
5	0	0	3	2	7	-7	25	1	0	10	5	15.5	10
7	0	0	4	2	8	-5	27	0	0	11	5	16.5	12
8	1	0	5	2	9	-3	34	1	0	12	5	17.5	14
<12	0	1	5	3	6	-9	<35	1	1	12	6	9.5	-2
15	0	0	6	3	10.5	0	40	1	0	13	6	19	17
17	1	0	7	3	11.5	2	43	1	0	14	6	20	19

- List the combined m background and n site measurements, including the less-than values, from smallest to largest, as illustrated in column 1 of Table 1. Ignore the less-than symbol when listing the N data from smallest to largest.
- Place a 0 or 1 in the second column of Table 2 (the column with heading h_i) using the following rule:
  h_i = 0 if the i^{th} measurement is from background
  = 1 if the i^{th} measurement is from the site
- Place a 0 or 1 in the 3rd column of Table 1 (the column with heading $INDEX_i$) using the following rule:
  $INDEX_i$ = 0 if the i^{th} measurement is a detect
  = 1 if the i^{th} measurement is a less-than values
- When moving down the data in column 1, determine the values of parameters d and e (columns 4 and 5 in Table 1) using the following rules:
  - If the first datum in column 1 is a detect, that is, if $Index_1$ = 0, then set d = 1 and e = 0 in the first row of Table 1.
  - If the first datum in column 1 is a less-than value, that is, if $Index_1$ = 1, then set d = 0 and e = 1 in the first row of Table 1.
  - For each successive row (rows 2 through N = 20) increase d by 1 whenever the datum in column 1 in that row is a detect, that is, whenever Index = 0
  - For each successive row increase e by 1 whenever the datum in column 1 in that row is a less-than value, that is, when index = 1.
- Let T denote the total number of less-than values in the pooled background and site data sets. For the previous

data there are T = 6 less-than values. Compute the rank of the i^{th} datum (i.e., of the datum in the i^{th} row in the previous table) as follows:

- $R_i = d_i + (T + e_i)/2$ if the datum in column 1 of the i^{th} row is a detect, that is, if $h_i = 0$ for the i^{th} row.
- $R_i = (T + 1 + d_I)/2$ if the datum in column 1 of i^{th} row is a less-than value, that is, if $h_i = 1$ for the i^{th} row.

- Compute the N = 20 scores, $a(R_1)$, $a(R_2)$, …, $a(R_{20})$, using the formula

  $$a(R_i) = 2R_i - N - 1$$

  for successive values of i = 1, 2, .., 20.

8. Compute the Gehan statistic, G:

$$G = \frac{(-11) + (-12) + (-3) + 2 + 6 + 10 + 14 + (-2) + 17 + 19}{\{10*10[(-13)^2 + (-11)^2 + (-12)^2 + \ldots + (-2)^2 + (17)^2 + (19)^2] / 20*19\}^{1/2}}$$

$$= 40 / [(100*1942) / (20*19)]^{1/2}$$
$$= 40 / 22.606$$
$$= 1.77$$

9. In Step 1 above we specified that $\alpha = 0.05$. Entering Table A.1with $\alpha = 0.05$ we find $Z_{1-\alpha} = Z_{0.95} = 1.645$. As G > 1.645, that is, 1.77 > 1.645, the Gehan test declares that the chemical is a COPC.

Box 3.19. Procedure for Conducting the Gehan Test when m and n are Less than 10.

1. Generate on a computer all possible orderings of the combined N site and background measurements. Denote the *number* of possible orderings by M.
2. Compute the G statistic (Equation 1 in Box 3.17) for each of these orderings to generate an empirical distribution (histogram) of M values of G.
3. Determine the $100(1-\alpha)^{th}$ percentile of the empirical distribution of G generated by Step 2 as follows (from Gilbert 1987, page 141) where α is the probability that can be tolerated of the test procedure described in this box incorrectly declaring that the chemical is a COPC:
   - Order the M values of G from smallest to largest.
   - Compute $k = (1-\alpha)(M + 1)$
   - If k is an integer, the $(1-\alpha)^{th}$ percentile is the k^{th} largest value of the ordered M values of G.
   - If k is not an integer, determine the value of k', where k' is the largest integer less than k. Compute the $(1-\alpha)^{th}$ percentile by linear interpolation between the k'th and $(k' +1)^{th}$ largest values of G.
4. If the value of G computed, *using the ordering actually observed for the collected background and site data*, equals or exceeds the $100(1 - \alpha)^{th}$ percentile obtained above, conclude the chemical is a COPC.

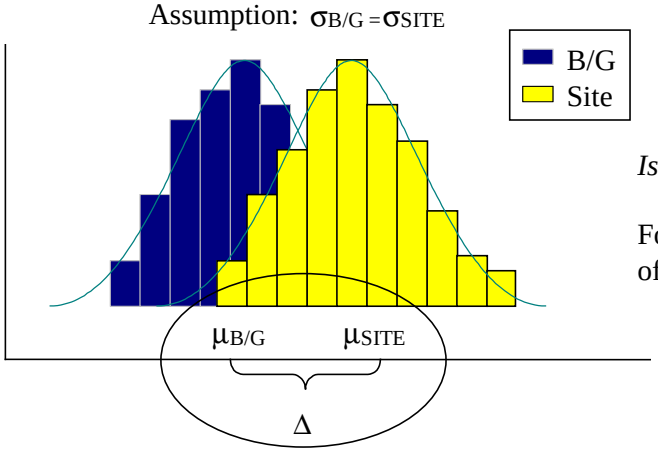
3.8 Two-Sample t Test

Site Contamination Scenario

The site contamination scenario for the two-sample t test is the same as that for the WRS test (Section 3.6). That is, if contamination from Navy site operations has occurred, it would probably be homogeneously distributed throughout the region.

Asks if the true mean of the site distribution is greater than the true mean of the background distribution.

Assumption: $\sigma_{B/G} = \sigma_{SITE}$



Is site > background?

Focus on mean of two distributions.

Role of the Data Quality Objectives Process

The stakeholders and regulators used the DQO planning process to agree:

- on the methods that will be used to collect, handle, prepare and measure the soil samples
- on the value of the design parameters for determining the number of site and background measurements needed (discussion following)
- that it is likely that very few less-than values will be reported by the laboratory
- that the measurements are likely to be normally distributed, but tests for normality of the measurements (Section 2.6) will be conducted to assess the validity of this assumption
- that if the tests for normality indicate the measurements are not normally distributed or the number of background and site measurements are not sufficiently large (both n and m greater than, say, 25 or 30) so the estimated site and background means are not approximately normally distributed, the WRS test and the Quantile test will be used in place of the two-sample t test
- that the measurements from the site are expected to have approximately the same total variance (among measurements) as the background measurements
- if a statistical test (an F test described in Iman and Conover 1983, page 275, and EPA (1996, Box 4.5-2, pages 4.5-2)) indicate the site and background measurements may not

have the same variance, but both data sets appear to be normally distributed, then the Satterthwaite two-sample t test will be used (Section 3.9) to test for differences in the site and background means.

The two-sample t test was selected because the assumptions of normality, equal variances for background and site data, and the absence of less-than values were expected to be valid. However, once the measurements are obtained, these assumptions will be evaluated by observation and by using statistical tests. If the site and background variances appear to be approximately equal but the data are not normally distributed, the WRS test may be used in place of the two-sample t test. If the two data sets are not normally distributed and have unequal variances, the Quantile and Slippage tests may be used. The assumptions that underlie the use of the two-sample t test are summarized in Box 3.1.

Advantages and Disadvantages of the Two-Sample t Test

- If less-than values should occur and if those values are replaced by substitute values such as the detection limit or one-half the detection limit, then the two-sample t test could be computed. However, the test would give biased and perhaps misleading results. If there is only one detection limit (for example, if all less-than values are < 10), and no more than about 40 percent of both the site and background data are less-than values, the recommendation in this situation is to replace the two-sample t test with the WRS test. This recommendation is correct even though the data may be normally distributed. The Quantile test may also be used in conjunction with the WRS test. If the less-than values take on multiple values (<10, <15, etc,) the Gehan test should be used in place of the WRS test.

- Most statistical software packages will compute the two-sample t test.

Guidance on Implementing the Two-Sample t Test

The number of site (n) and background (m) measurements required to conduct the two-sample t test should be approximated using the procedure outlined in Box 3.20. An example of the computation of Equation 1 in Box 3.20 is given in Box 3.21. After n and m have been determined, the samples collected, and measurements reported by the laboratory, summary statistics (Section 2.3) should be computed for both the site and background data sets. In particular, the computed sample variance of the site measurements should be compared with the computed sample variance of the background measurements to determine if they are approximately equal, a required assumption of the two-sample t test. A procedure (an F test) for testing statistically if the two sample variances are equal is provided in Iman and Conover (1983, page 275). This procedure is commonly found in statistical software packages.

If some measurements appear to be unusually large compared to the remainder of the measurements in the data set, a test for outliers (Section 2.4) should be conducted. Once any identified outliers have been investigated for being mistakes or errors and, if necessary, discarded, the site and background data sets should be tested for normality using both probability plots (Section 2.5.4) and formal tests of hypotheses (Section 2.6).

After the assumptions of equal variances and normality have been shown to be reasonable, the two-sample t test can be conducted. The procedure for doing the test is given in Box 3.22. An example of the procedure is given in Box 3.23

Box 3.20. Procedure for Calculating the Number of Site and Background Measurements Required to Conduct the Two-Sample t Test

The formula for calculating the number of site (n) and background (m) measurements required to conduct the two-sample t test is:

$$n = m \approx \frac{2\sigma^2 (Z_{1-\alpha} + Z_{1-\beta})^2}{(\mu_s - \mu_b)^2} + 0.5*(Z_{1-\alpha})^2 \qquad (1)$$

where

σ^2 = expected variance of the measurements at both the site and background area (ideally, the value of σ^2 used should be approximated using measurements previous obtained from the site and background or obtained in a special pilot study at the site and background)

α = the probability that can be tolerated that the two-sample t test will incorrectly declare the chemical is a COPC (α is usually specified to be a small value such as 0.01, 0.025, 0.05 or 0.10)

$1 - \beta$ = the power (probability) required that the two-sample t test will declare the chemical is a COPC when that is indeed the case (β is usually specified to be ≥ 0.80)

$\mu_s - \mu_b$ = true site mean (μ_s) minus the true background mean (μ_b)
= the difference in the *true* (unknown) means of the site and background that the stakeholders and regulators have agreed needs to be detected by the two-sample t test with power (probability) equal to1 - β.

$Z_{1-\alpha}$ = the 100(1-α) percentile of the standard normal distribution, which is found in Table A.1 (for example, if $\alpha = 0.05$, Table A.1 indicates $Z_{1-0.05} = Z_{0.95} = 1.645$)

$Z_{1-\beta}$ = the 100(1-β) percentile of the standard normal distribution, which is found in Table A.1 (for example, if $1 - \beta = 0.80$, then we find from Table A.1 that $Z_{0.80} = 0.84$)

The appropriate values of the parameters in Equation 1 in this box should be determined by the stakeholders and regulators during the application of the DQO planning process.

Box 3.21. Example of the Procedure for Calculating the Number of Site and Background Measurements Required to Conduct the Two-Sample t Test

Suppose the values of the parameters in Equation 1 in Box 3.20 were specified by the stakeholders and regulators as follows:

$$\sigma^2 = 7.5$$
$$\alpha = 0.025$$
$$1-\beta = 0.80$$
$$\mu_s - \mu_b = 4$$

Looking in Table A.1 we find that

$$Z_{1-\alpha} = Z_{0.975} = 1.96 \quad \text{and}$$
$$Z_{1-\beta} = Z_{0.80} = 0.84$$

Hence, Equation 1 is:

$$n = m \approx 2*7.5*(1.96 + 0.84)^2 / 4^2 + 0.50*(1.96)^2$$
$$= 9.27 \text{ or } 10$$

Therefore, 10 site and 10 background measurements are required for the two-sample t test to attain the performance specified (by the values of α and $1-\beta$) to detect a difference in true means of size $\mu_s - \mu_b = 4$ when the variance of the data at the site and background areas is $\sigma^2 = 7.5$.

The reader may want to try other values of σ^2 and $\mu_s - \mu_b$ to see how n = m change for the specific values of α and $1 - \beta$ given above.

Box 3.22. Procedure for Conducting the Two-Sample t Test

1. Stakeholders and regulators use the DQO process to select values of σ^2, α, $1-\beta$ and $\mu_s - \mu_b$ and use the procedure in Box 3.20, as illustrated in Box 3.21, to determine the number of site (n) and background (m) measurements.
2. Collect the samples and obtain the n and m site and background measurements.
3. Suppose
   - the n site measurements are denoted by $x_1, x_2, \ldots, x_n$
   - the m background measurements are denoted by $y_1, y_2, \ldots, y_m$.
4. Compute the two-sample t test statistic, denoted by T:

$$T = \frac{\bar{x} - \bar{y}}{\{(n+m)[n-1)s_x^2 + (m-1)s_y^2] / [nm(n+m-2)]\}^{1/2}}$$

where $\bar{x}$ = the arithmetic mean of the n site measurements

$\bar{y}$ = the arithmetic mean of the m background measurements

s_x^2 = the sample variance of the n site measurements (the formula for computing s_x^2 is given in the 8th row of Box 2.1 in Section 2.3.1)

s_y^2 = the sample variance of the m background measurements (see Box 2.1)

5. The two-sample t test declares:

   - the chemical is a COPC if $T \geq t_{1-\alpha,\, n+m-2}$
   - insufficient evidence exists to conclude that the chemical is a COPC if $T < t_{1-a,\, n+m-2}$

where $t_{1-\alpha,\, n+m-2}$ is the $100(1-\alpha)$ percentile of the t distribution that has n + m - 2 degrees of freedom. The value of $t_{1-\alpha,\, n+m-2}$ is determined from Table A.16 by entering that table with the values of $1 - \alpha$ and n + m -2. Note the value of α was specified in Step 1, as part of the process for determining the number of site and background measurements required.

If the two-sample t test declares the chemical is not a COPC, it may indicate (1) the chemical is indeed not a COPC, or (2) the assumptions that underlie the t test are not valid for the site and background measurements, or (3) an insufficient number of measurements (n and m) were obtained for the t test to be able to detect the difference in site and background concentration distributions that actually exists. An evaluation should be made of the possibility the causes in items 2 or 3 may have resulted in the t test declaring the chemical is not a COPC.

- First, review the DQO planning process records to make sure the number of samples (n and m) collected agree with what was determined at that time to be necessary to detect a difference between site and background means that was considered important.
- Second, review the computations conducted to test for normality and equality before the t test was calculated. Verify that the tests were done correctly using the appropriate data. Redo the tests if necessary.
- Third, the shift in the site concentration distribution may in fact be smaller than the shift selected by the stakeholders as being important to detect, in which case no additional measurements are needed. However, as the true difference in means is unknown, update the estimated number of site and background measurements needed to detect the critical (important) shift in the site mean by calculating the variance of the site and background measurements (s_x^2 and s_y^2, respectively) and use the larger of these two estimated variances in Equation 1 of Box 3.20. If this new value, denoted by n′, is larger than either the number of site or background measurements obtained and used in the t test, collect additional samples so n′ site and n′ background measurements are available. Then redo the t test.

Box 3.23. Example of Computations for the Two-Sample t Test

1. Suppose the values of the parameters in Equation 1in Box 3.20 were specified by the stakeholders and regulators to be $\sigma^2 = 7.5$, $\alpha = 0.025$, $1-\beta = 0.80$, and $\mu_s - \mu_b = 4$. In Box 3.21 it was shown that $n = m = 10$ for these parameter values.
2. The n = m measurements were obtained.
3. Suppose the values were as follows

   Site Measurements (x) : 90, 77, 81, 210, 92, 130, 110, 120, 140, 84
   Background Measurements (y) : 23, 15, 78, 26, 90, 99, 87, 34, 17, 10

   There do not appear to be any potential outliers in either data set. Hence, tests for outliers (Section 2.3) do not appear to be needed. Each data set should be used in a test for normality (Section 2.6). The reader is encouraged to conduct these tests. Suppose the tests indicate the data can be assumed to be normally distributed.
4. The following calculations were conducted:

$$\bar{x} = 113.4$$

$$\bar{y} = 47.9$$

$$s_x^2 = 1623.82$$

$$s_y^2 = 1287.21$$

$$T = \frac{\bar{x} - \bar{y}}{\{ (n + m)[(n-1) s_x^2 + (m-1) s_y^2] / [nm (n + m - 2)] \}^{1/2}}$$

$$= \frac{113.4 - 47.9}{\{ (10 + 10)[9*1623.82 + 9*1287.21] / [10*10 (10 + 10 - 2)] \}^{1/2}}$$

$$= \frac{65.5}{17.06}$$

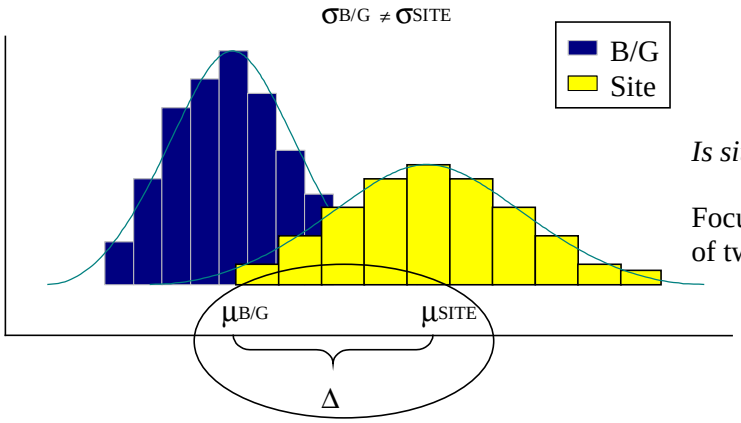
$$= 3.84$$

5. The value of $t_{1-\alpha, n+m-2}$, that is, of $t_{0.975, 18}$ is found from Table A.16 to be 2.101. Hence, as T > 2.101, that is, 3.84 > 2.101, the two-sample t test declares that the chemical is a COPC.

3.9 Satterthwaite Two-Sample t Test

Site Contamination Scenario

Asks if the true mean of the site distribution is greater than the true mean of the background distribution.



Is site > background?

Focus on mean of two distributions

The site contamination scenario for the Satterthwaite two-sample t test is the same as that for the usual two-sample t test (Section 3.8). That is, if contamination from Navy site operations has occurred, it would probably be homogeneously distributed throughout the region.

Data Quality Objectives

The only difference between the DQOs for the usual two-sample t test (Section 3.8) and the Satterthwaite two-sample t test is that the stakeholders and regulators have concluded, based on prior data and statistical tests or on the basis of expert knowledge, that the measurements from the site are *not* expected to have approximately the same total variance (among measurements) as the background measurements. Recall from Section 3.8 that a procedure for testing statistically if two sample variances are equal is provided in Iman and Conover (1983, page 275) and EPA (1996, Box 4.5-2, pages 4.5-2)).

Limitations and Robustness (Advantages/Disadvantages) of the Satterthwaite Two-Sample t Test

If less-than values should occur and if those values are replaced by substitute values, such as the detection limit or one-half the detection limit, then the Satterthwaite two-sample t test could be computed. However, the test would give biased and perhaps misleading results. The recommendation in this situation is to replace the Satterthwaite t test with the WRS and Quantile tests. If the less-than values take on multiple values, (for example, <10, <15, etc,) the Gehan test should be used in place of the WRS test.

Guidance on Implementing the Satterthwaite Two-Sample t Test

It is recommended the same number of measurements should be obtained for both the site and background areas. Let the number of such measurements be denoted by n. The number of site and background measurements should be approximated using the procedure in Box 3.20 that was used for the two-sample t test, *where s^2 in Equation (1) of Box 3.20 is now the larger of the site and background measurement variances.*

When the n measurements have been obtained and the assumption of normality appears reasonable based on the use of graphical methods (Section 2.5) and statistical tests (Section 2.6), the Satterthwaite test can be conducted as described in Box 3.24. An example of the Satterthwaite test is given in Box 3.25.

Box 3.24. Procedure for Conducting the Satterthwaite Two-Sample t Test

1. Use the DQO process to select values of α, β, $\mu_s - \mu_b$ and the larger of the site and background variances (σ^2). Then use the procedure in Box 3.20, as illustrated in Box 3.21 to determine the number of measurements (n) for both the site and the background area.
2. Collect the samples and obtain the n site and n background measurements
3. Suppose
   - the n site measurements are denoted by $x_1, x_2, \ldots, x_n$
   - the n background measurements are denoted by $y_1, y_2, \ldots, y_n$
4. Compute the Satterthwaite two-sample t test statistic, denoted by T_s:

$$T_s = \frac{\bar{x} - \bar{y}}{\left(s_x^2/n + s_y^2/n\right)^{1/2}}$$

where $\bar{x}$ = the arithmetic mean of the n site measurements
$\bar{y}$ = the arithmetic mean of the n background measurements
s_x^2 = the sample variance of the n site measurements (the formula for computing is given in the 8th row of Box 2.1)
s_y^2 = the sample variance of the n background measurements (see Box 2.1).

5. Compute the approximate degrees of freedom, *f*, as follows:

$$f = \frac{\left(s_x^2/n + s_y^2/n\right)^2}{\left(s_x^2/n\right)^2/(n-1) + \left(s_y^2/n\right)^2/(n-1)}$$

Note: the Satterthwaite t-test can be computed when the number of site and background measurements are not equal. In that case, n in these equations would be replaced by n_x and n_y, as appropriate.

6. The Satterwaite two-sample t test declares that:

- the chemical is a COPC if $T_s \geq t_{1-\alpha, f}$
- insufficient evidence is offered to conclude that the chemical is a COPC if $T_s < t_{1-\alpha, f}$, where $t_{1-\alpha, f}$ is the 100(1 - α) percentile of the t distribution that has f degrees of freedom. The value of $t_{1-\alpha, f}$ is determined from Table A.16 by entering that table with the values of 1 - α and f. Linear interpolation may be used to determine $t_{1-a, f}$ in Table A.16 if f is not an integer.

If the two-sample t test declares the chemical is *not* a COPC, it may indicate (1) the chemical is indeed not a COPC, or (2) the assumptions that underlie the t test are not valid for the site and background measurements, or (3) an insufficient number of measurements (n and m) were obtained for the Satterthwaite t test to be able to detect the difference in site and background concentration distributions that actually exists.

An evaluation should be made of the possibility the causes in items 2 or 3 may have resulted in the t test declaring that the chemical is not a COPC.

- First, review the DQO planning process records to make sure that the number of samples (n and m) collected agrees with what was determined at that time to be necessary to detect a possible difference between site and background means that was considered important.
- Second, review the computations done for the tests for normality and equality of variance conducted on the measurements before the Satterthwaite t test was calculated. Verify the tests were done correctly using the appropriate data. Redo the Satterthwaite t tests if necessary.
- Third, the shift in the concentration distribution may, in fact, be smaller than the shift selected by the stakeholders, in which case no additional measurements are needed. However, as the true difference in means is unknown, update the estimated number of site and background measurements needed by calculating the variance of the site and background measurements (s_x^2 and s_y^2, respectively) and use the larger of these two estimated variances in Equation 1 of Box 3.20. If this new value, denoted by n', is larger than the number of site and background measurements obtained and used in the t test, then collect additional samples so that n' site and n' background measurements are collected. Then redo the Satterthwaite t test.

Box 3.25. Example of the Procedure for Conducting the Satterthwaite Two-Sample t Test

1. Suppose a preliminary study was conducted to estimate the variance of the background and site measurements and the variance of the site data was significantly larger than the background data variance. Suppose the larger of the two estimated variances was 15. Hence, that value was selected as the value for σ^2 to use in Equation 1 in Box 3.20. (If very few site and background measurements were obtained in the preliminary study, say less than 10 for each, the value for σ^2 may be increased by 20% or so to guard against not taking enough measurements.) Also, suppose the values of the other parameters in Equation 1 in Box 3.20 were specified by the stakeholders and regulators during the DQO process to be $\alpha = 0.10$, $1 - \beta = 0.90$ and $\mu_s - \mu_b = 4$. For these parameter values, the reader may verify that Equation 1 in Box 3.20 gives the value n = 13.1, rounded up to n = 14.
2. Therefore, n = 14 site and n = 14 background samples were collected and measured using the methods specified during the DQO process and as documented in the Quality Assurance Project Plan (QAPP).
3. Suppose the measurements are as follows:

   Site Measurements (x) : 7.2, 3.3 10.9, 11.5, 2.0, 6.4, 12.1, 2.2, 0.5, 0.9, 1.1, 2.0, 5.1, 10.5
   Background Measurements(y): 8.1, 13.2, 5.0, 2.5, 7.2, 3.9, 10.8, 1.1, 8.5, 11.3, 9.2, 2.7, 3.1, 9.1

4. No potential outliers appear to be present in either data set. Hence, tests for outliers (Section 2.4) do not appear to be needed. Each data set should be evaluated graphically (Section 2.5) and using a formal statistical test (Section 2.6) to evaluate if the data for each data set can be reasonably assumed to be normally distributed. The reader may verify the assumption of normality appears to be a reasonable assumption for both data sets.
5. Next, the following calculations are conducted:

$\bar{x} = 5.41$

$\bar{y} = 6.84$

$s_x^2 = 18.708$

$s_y^2 = 14.316$

$$T_s = \frac{\bar{x} - \bar{y}}{(s_x^2/n + s_y^2/n)^{1/2}}$$

$$= \frac{5.41 - 6.84}{(18.708/14 + 14.316/14)^{1/2}}$$

$$= \frac{-1.43}{1.536}$$

$$= -0.931$$

$$f = \frac{(18.708/14 + 14.316/14)^2}{(18.709/14)^2/13 + (14.316/14)^2/13}$$

$$= \frac{5.564}{0.1374 + 0.08043}$$

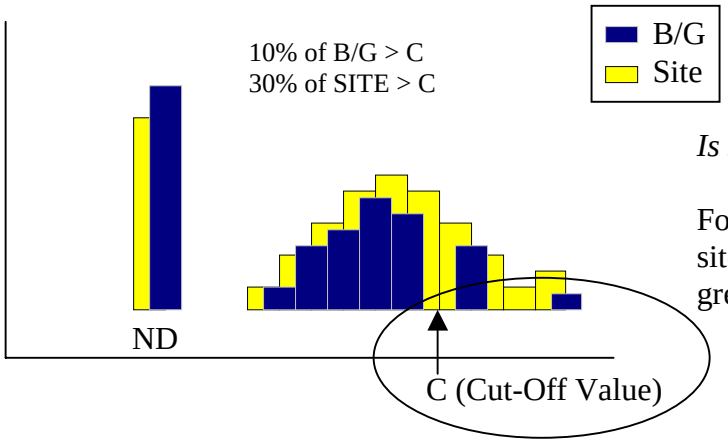
= 25.54 degrees of freedom

6. Using linear interpolation between $t_{0.90,25} = 1.316$ and $t_{0.90,26} = 1.315$ in Table A.16, we find that $t_{0.90,25.54} = 1.3155$. Hence, as $T_s < 1.3155$, that is, as $-0.931 < 1.3155$, the Satterthwaite two-sample t test does *not* declare the chemical is a COPC. Indeed, the estimated mean of the site measurements is less than the estimated mean of the background measurements.
7. As the test did not declare the chemical was a COPC, the DQO process records and QAPP should be reviewed to double check that all requirements for collecting the type, quantity, and quality of measurements were correctly followed. Next, evaluate whether the number of measurements used in the test (n = 14) was too small to achieve the allowable α and β decision error rates specified during the DQO process (see Step 1 in this example) and recorded in the QAPP. To do so, we compute Equation 1 in Box 3.20 using the larger of the estimated site and background variances, that is, using $\sigma^2 = 18.7$, as computed in Step 5. We find from Equation 1 that n = 16.1, which is rounded up to 17, when $\sigma^2 = 18.7$, $\alpha = 0.10$, $1 - \beta = 0.90$ and $\mu_s - \mu_b = 4$. Hence, 3 additional samples should be collected and measured in both the background area and at the site. Simple random sampling should be used to determine the locations in the field of the new samples. Also, the collection and measurement protocols specified in the QAPP for obtaining the new data should be exactly the same as for the original data. Then the Satterthwaite two-sample t test should be recomputed using the new background and site data sets, each of which consists of 14 old and 3 new measurements. Before conducting the Satterthwaite t test the graphical methods and statistical test for normality should be conducted on the new data sets (n = 17) to reassess if the normality assumption is still reasonable.

3.10 Two-Sample Test of Proportions

Site Contamination Scenario

Asks if a larger proportion of the site data than of the background data exceeds a concentration C.



Is site > background?

Focus on proportion of site and background data greater than C.

Suppose Navy operations may have released a contaminant to surface soil in a region (Region A) on the Navy site, but a distinct contamination pattern of high and low concentrations is not expected to be present. Therefore, a statistical test will be applied to the entire Region A to indicate whether to reject the null hypotheses indicating the chemical of interest is not a COPC and accept the null hypothesis that the chemical of interest *is* a COPC. If a distinct contamination pattern were expected to have occurred, and if sufficient information on that pattern was available or could be obtained, Region A would be separated into separate strata (subregions) that are relatively homogeneous. In that case, a separate statistical test and decision would be made for each stratum.

Data Quality Objectives

The DQO planning process was implemented. Suppose the DQO planning team, including regulators, believed it was highly likely that more than 50% of the background, and possibly site, measurements would be reported as less-than values. In this case, it is difficult to conduct a valid statistical test of whether the site average (mean or median) is shifted to the right (to higher concentrations) of the background average (mean or median). Therefore, the DQO planning team decided to conduct a statistical test to assess if a larger *proportion* of the site than of the background area had concentrations greater than a specified concentration C, where C is greater than the detection limit. The two-sample test for proportions is suitable for this situation.

The DQO planning team also agreed:

- that the null and alternative hypotheses that will be tested are

  H_o: $P_s \leq P_b$

  H_a: $P_s > P_b$

where P_s and P_b are the true proportions of the site and background distributions of potential measurements, respectively, that exceed C. When this H_o and H_a are used, the burden of proof is on showing that $P_s > P_b$.

- on the methods that will be used to collect, handle, prepare, and measure the soil samples

- that the value of the concentration C should be just slightly greater than the largest background less-than value and, therefore, C would need to be selected after the background data are obtained

- on the parameters needed to compute the number of background and site measurements of the chemical of interest (discussed following).

The assumptions that underlie the use of the two-sample test of proportions are summarized in Box 3.1

Limitations and Robustness (Advantages/Disadvantages) of the Two-Sample Test for Proportions

- The test may be conducted regardless of the underlying distribution of the measurements. That is, the test is a distribution-free (non-parametric) test.

- The test is rather easy to perform.

- However, the test requires that the measurements be independent (not spatially or temporally correlated) and that simple random sampling be used to determine the sampling locations in both the background and site areas. However, sampling on a grid pattern is acceptable if the grid pattern does not correspond (line up) with a pattern of changing concentrations for the chemical of interest in either the background or site areas.

- The test does not test whether the site mean (median) exceeds the background mean (median).

The two-sample proportion test is being used in this case because it is not possible to avoid a large number of less-than values when the measurement method of choice is used.

Guidance on Implementing the Two-Sample Test for Proportions

The number of site (n) and background (m) measurements required to conduct the two-sample test for proportions should be approximated using the procedure outlined in Box 3.26. An example of the procedure is given in Box 3.27. When the data are collected according to the specifications worked out during the DQO planning process, the data sets should be examined to look for outliers. A test for outliers should be conducted for any datum that appears to be unusually large, relative to the remaining data in the data set. Tests for normality or lognormality (Sections 2.5 and 2.6) of the data need not be conducted.

After the DQA process has been completed, (that is, once it has been determined that the data contain no errors, that they have been collected, handled, and measured according to the specifications developed during the DQO process), and that the assumptions that underlie the use of the two-sample test for proportions have been shown to be reasonable, then the test may be conducted. The procedure for conducting the test is given in Box 3.28. An example is provided in Box 3.29.

Box 3.26. Procedure for Calculating the Number of Site and Background Measurements Required to Conduct the Two-Sample Test for Proportions

The formula for calculating the number of site (n) and background (m) measurements required to conduct the two-sample test for proportions is as follows (from EPA 1996, page 3.3-8):

$$n = m = \frac{2(Z_{1-\alpha} + Z_{1-\beta})^2 \ \overline{P}\ (1- \overline{P}\ )}{D^2} \qquad (1)$$

where

$\overline{P}$ = $(P_s + P_b)/2$

P_s = the proportion of the true site distribution of potential measurements that exceeds C

P_b = the proportion of the true background distribution of potential measurements that exceeds C.

α = the probability that can be tolerated that the two-sample test for proportions will incorrectly reject H_o, that is, will incorrectly declare the chemical is a COPC, (α is usually specified to be a small value such as 0.01, 0.025, 0.05 or 0.10)

$1 - \beta$ = the power (probability) required that the two-sample test for proportions will declare that the chemical is a COPC when that is indeed the case, (β is usually specified to be $\geq$ 0.80)

D = the difference in the true (unknown) proportions of the site and background distributions of potential measurements that exceed the constant C *that must be detected with probability 1 - b*. That is, the stakeholders and regulators have agreed that the difference D needs to be detected by the two-sample test for proportions with power (probability) equal to $1 - \beta$.

$Z_{1-\alpha}$ = the $100(1-\alpha)$ percentile of the standard normal distribution, that is tabulated in Table A-1 (for example, if $\alpha = 0.05$, then Table A-1 indicates that $Z_{1-0.05} = Z_{0.95} = 1.645$)

$Z_{1-\beta}$ = the $100(1-\beta)$ percentile of the standard normal distribution, that is tabulated in Table A-1 (for example, if $1 - \beta = 0.80$, we find from Table A-1 that $Z_{0.80} = 0.84$)

The appropriate values of the parameters in Equation 1 should be determined by the stakeholders and regulators during the application of the DQO planning process.

Box 3.27. Example of the Procedure for Calculating the Number of Site and Background Measurements Required to Conduct the Two-Sample Test for Proportion

Suppose the values of the parameters in Equation (1) of Box 3.26 were specified by the stakeholders and regulators as follows:

$D = 0.20$
$\alpha = 0.025$
$\beta = 0.20$
$Z_{1-\alpha} = Z_{0.975} = 1.96$ and $Z_{1-\beta} = Z_{0.80} = 0.84$ (from Table A-1).

Since P_s and P_b are true values and hence are unknown, estimates of these true proportions must be supplied from a preliminary sampling study conducted at the background and site. This study must be conducted using the same sampling and analysis protocol that will be used in the main study. Suppose a preliminary study based on collecting 20 samples in the background area and 20 samples at the site yields estimates of P_s and P_b to be 0.30 and 0.15, respectively. Hence, $\overline{P} = (0.30 + 0.15) / 2 = 0.225$. Hence, equation (1) in Box 3.26 is:

$$n = m = 2(1.96 + 0.84)^2\, 0.225(1 - 0.225) / 0.20^2$$
$$= 68.35$$

and rounded up to 69. Hence, 69 samples are needed from the background area and 69 from the site. Because the 20 site and 20 background samples have already been collected, handled, and measured using the methods required for the full study, only 49 new site and 49 new background measurements need be collected.

Box 3.28. Procedure for Conducting the Two-Sample Test for Proportions (from EPA 1998)

1. Stakeholders and regulators use the DQO process to select values of α, β, D and C (recall that C is the concentration limit of interest; see the Data Quality Objectives section at the beginning of Section 3.10).
2. Conduct a preliminary sampling and measurement study at the background area and for the region within the Navy site being examined (Region A) to obtain estimates of the true proportions P_s and P_b of the site and background populations that exceed C. Then use the procedure in Box 3.26 to determine n and m, the number of site and background measurements needed.
3. Collect, handle, and measure the n and m samples, as specified in the sampling and analysis plan and the QAAP.
4. Suppose
   - n site measurements are denoted by $x_1, x_2, \ldots, x_n$
   - m background measurements are denoted by $y_1, y_2, \ldots, y_m$

   Note: In this handbook we recommend that n = m. However, the following formulas are for the more general case where the number of site measurements, n, and the number of background measurements, m , are not equal.
5. Let k_s and k_b be the number of site and background measurements, respectively, that exceed C.
6. Compute $p_s = k_s / n$, which is the estimated proportion of the true distribution of potential site measurements that exceed C.
7. Compute $p_b = k_b / m$, which is the estimated proportion of the true distribution of potential background measurements that exceed C.
8. Compute
   $$p = (k_s + k_b) / (n + m)$$
9. Compute np_s , mp_b, $n(1-p_s)$, $m(1-p_b)$. If all of these quantities are greater than or equal to 5, continue with step 10. If not, seek assistance from a statistician as the computations for the test become more complicated.
10. Compute the test statistic:
    $$Z_p = (p_s - p_b) / [p(1- p)(1/n + 1/m)]^{1/2}$$
11. Use Table A-1 in this handbook to find $Z_{1-\alpha}$
12. If $Z_p \geq Z_{1-\alpha}$ the test has declared that $P_s > P_b$, that is, that the true proportion of the potential site measurements greater than the concentration value C is greater than the true proportion of the potential background measurements greater than C.

    If $Z_p < Z_{1-\alpha}$ then not enough evidence is present from the data to conclude that $P_s > P_b$. In that case, go to step 13.
13. Suppose the test declares not enough evidence is present from the data to conclude that $P_s > P_b$. This conclusion may indicate (1) the chemical is not a COPC, or (2) the assumptions that underlie the test are not valid for the site and background measurements, or (3) an insufficient number of measurements (n and m) were obtained for the test to be able to detect the difference D that actually exists. Evaluate if the causes in items 2 or 3 may have resulted in the test declaring the chemical is not a COPC. Review the DQO planning process records to make sure the number of measurements (n and m) agree with what was determined at that time to be necessary to detect the specified difference D. For item 3, use equation (1) in Box 3.26 to recompute the number of measurements required for the test. Those computations should be done using the estimates p_s and p_b in place of P_s and P_b , respectively. If the new value of n is greater than what was used to compute the test statistic, collect the additional samples needed and redo the test.

Box 3.29. Example of Computations for the Two-Sample Test for Proportions

1. Suppose the stakeholders and regulators specified that $\alpha = 0.025$, $\beta = 0.20$, D = 0.20 and C = 1 ppb for the chemical of interest.
2. Also suppose that a preliminary study was conducted at the site and background area to obtain estimates of the true proportions P_s and P_b. Suppose these estimates were 0.30 and 0.15, respectively. Then, as illustrated in Box 3.27, n = 69 measurements are needed from the site and 69 also from the background area.
3. A total of 138 measurements are obtained. Suppose $k_b = 19$ of the 69 background measurements were greater than C, that is, greater than 1 ppb. Furthermore, suppose that $k_s = 24$ of the site measurements were greater than C. Hence,
   $p_b = 19/69 = 0.275$
   $p_s = 24/69 = 0.347$
   $p = (k_s + k_b) / (n + m) = (19 + 24) / (69 + 69) = 0.3116$
4. Also,
   $mp_b = 69*0.275 = 19$
   $mp_s = 69*0.347 = 24$
   $m(1-p_b) = 69(1-0.275) = 50$
   $n(1-p_s) = 69*(1-0.347) = 45$
   all greater than 5. Hence, we continue on with the test as described in Box 3.28.
5. The test statistic is computed as follows:

$$Z_p = (p_s - p_b) / [p(1- p)(1/n + 1/m)]^{1/2}$$

$$= (0.347 - 0.275) / [0.3116(1 - 0.3116)(1/69 + 1/69)]^{1/2}$$
$$= 0.072 / [0.2145*(0.014493 + 0.014493)]^{1/2}$$
$$= 0.072 / 0.0789$$
$$= 0.913$$

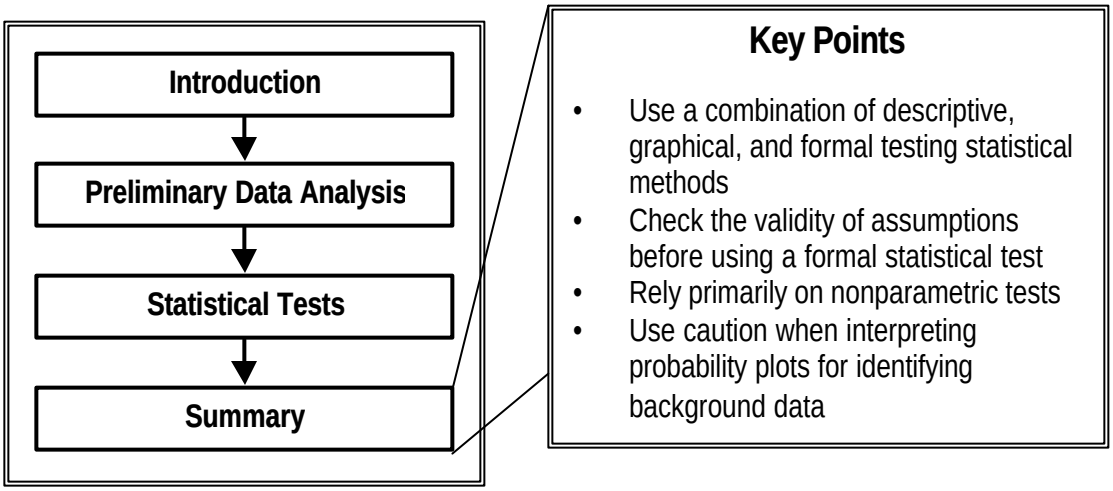
6. From Table A-1 we find that $Z_{1-\alpha} = Z_{0.975} = 1.96$
7. As $Z_p < 1.96$, that is, $0.913 < 1.96$, the data do not provide sufficient information for the test to reject H_o and declare the chemical is a COPC.
8. We re-compute Equation (1) in Box 3.26 to check if this lack of a statistically significant results (at the $1 - \alpha = 0.975$ confidence level) is due to collecting fewer measurements than required to achieve the power of $1 - \beta = 0.80$ when D = 0.20. We obtain:

$$n = m = 2(1.96 + 0.84)^2 \ 0.3116(1 - 0.3116) /0.2^2 = 84.09$$

that indicates 85 site and 85 background measurements are needed. Hence, too few measurements have been made.

9. In conclusion, the data indicate the true difference D is estimated to be 0.347 – 0.275 = 0.072. The two-sample test for proportions was not able to declare on the basis of the data this difference was large enough to conclude that $P_s > P_b$.

4.0 SUMMARY



This handbook provides detailed instructions for computing descriptive statistics and conducting graphical and statistical analyses to determine if concentrations of chemicals in soil at Navy sites contributed by Navy operations are significantly elevated relative to concentrations in ambient (local) background areas. If so, the chemicals are declared to be contaminants of potential concern (COPC) and they will be carried forward into human and ecological risk assessments. Uncertainty in these decisions that arises from being able to collect and measure only a limited number of soil samples (due to inevitable resource constraints) is taken into account by the use of statistical tests of hypotheses.

The key questions that are addressed in this handbook are:

- When can two or more data sets be combined to improve the chances of detecting when a chemical is a COPC?
  - See Section 2.2.
- How can we statistically describe and graphically explore background and Navy site data sets to look for differences in chemical concentrations and to assess which statistical decision rules (tests of hypotheses) should be used to identify COPC?
  - See Sections 2.3 (Descriptive Summary Statistics), 2.4 (Determining Presence of Data Outliers) and 2.5 (Graphical Data Analyses).
- What statistical procedures or tests should we use to determine if a chemical is a COPC?
  - See Section 3.1.1 (Selecting a Statistical Test).
- What are two decision rules that should be avoided in order to not falsely conclude a chemical is a COPC?
  - See Section 3.3.
- How do I determine the number of samples to take and how do I perform the selected statistical tests?

- See Sections 3.4 through 3.10.

Following are some general words of advice about using statistical methods when deciding which chemicals are COPC:

- Use a combination of descriptive statistics, graphical methods and formal tests of hypotheses.
- Use graphical probability plotting methods only as an initial step in determining COPC. Also use descriptive statistics, other graphical plots of the data, information on the natural correlation among chemicals in soil that may exist in nature, and statistical tests for COPC.
- Always check the assumptions that underlie a formal statistical test of hypothesis for COPC. For example, some statistical tests require that the data be normally distributed or that the variances of the site and background data be equal.
- Use the nonparametric Slippage test (Section 3.4) (comparing site measurements to the maximum background measurement) as a quick way to test for COPC.
- Use the Quantile test (Section 3.5) if an important criteria for deciding which chemicals are COPC is whether the right tail of the site concentration distribution is shifted to higher values than the right tail of the background concentration distribution.
- Consider using the nonparametric WRS test to decide if a chemical is a COPC if the assumptions that underlie the two-sample t test or the Satterthwaite two-sample t test are unreasonable for the Navy site of interest. See Section 3.6.
- Use the Gehan test instead of the WRS test if the background or site data sets contain multiple less-than values. See Section 3.7.
- Use the two-sample t test if the background and site data are normally distributed, if both data sets have about the same variance among the data, and if very few or no less-than values are present. See Section 3.8.
- Use the Satterthwaite two-sample t test if the background and site data sets are normally distributed, one data set has a larger variance than the other data set, and very few less-than values are present. See Section 3.9.
- Use the two-sample test for proportions if more than 50 percent of the background or site measurements are less-than values. See Section 3.10.
- Avoid comparing the maximum background measurement with the maximum site measurement to decide if a chemical is a COPC (Section 3.3.1). Similarly, avoid comparing the maximum site measurement to a background threshold value (Section 3.3.2). These decision rules can have a high probability of falsely concluding that the chemical is a COPC.
- Expect to use nonparametric tests most of the time (Slippage, Quantile, WRS, Gehan, and the two-sample test of proportions) because they allow for the occurrence of more less-than values and for arbitrarily shaped data sets.
- Consult an experienced environmental statistician whenever questions arise regarding the most appropriate graphical or statistical testing methods to use. The application of statistics requires a thorough knowledge of statistical methods for environmental applications and the conditions for which they should be used.

5.0 GLOSSARY

Alternative Hypothesis, H_a	The hypothesis that is accepted if the null hypothesis is rejected.
α	Alpha is the probability tolerated of falsely rejecting the null hypothesis and accepting the alternative hypothesis. Alpha is specified in Step 6 of the DQO process.
β	Beta is the probability tolerated of falsely accepting the null hypothesis as being true. Beta is specified in Step 6 of the DQO process.
Censored Data Set	A censored data set is one that contains one or more non-detects.
DQA Process	The Data Quality Assessment (DQA) process is the scientific and statistical evaluation of data to determine if data obtained from environmental data operations are of the right type, quality, and quantity to support their intended use.
DQO Process	The DQO process is a series of planning steps based on the scientific method designed to ensure the type, quantity, and quality of environmental data used in decision-making are appropriate for the intended application.
Less-than Values	Less-than values are non-detects that are reported by the analytical laboratory as being less than some quantitative upper limit value, such as the detection limit or the quantitation limit.
Non-detects	Non-detects are measurements that are reported by the analytical laboratory to be below some quantitative upper limit, such as the detection limit or the quantitation limit. A non-detect has insufficient measurement certainty for the analytical laboratory to report the chemical being measured is assured to be present at a quantifiable level in the sample.
Null Hypothesis, H_o	The hypothesis that is assumed to be true, unless the data indicate with sufficient confidence that it should be rejected in favor of the alternative hypothesis, H_a.
Power	Power is the probability the null hypothesis is rejected, when it is indeed false. Power is defined to be $1 - \beta$.
Target Population	The set of environmental space/time units within spatial and time boundaries for which a decision is needed as to whether a chemical of interest is a COPC.

6.0 REFERENCES

Akritas, M.G., T.F. Ruscitti, and G.P. Patil. 1994. Statistical Analysis of Censored Environmental Data, pp. 221-242, *Handbook of Statistics, Volume 12, Environmental Statistics*, (G.P. Patil and C.R. Rao, editors), North-Holland, New York.

Conover, W.J. 1980. *Practical Nonparametric Statistics*, 2nd edition, Wiley, New York.

D'Agostino, R.B. 1971. An Omnibus Test of Normality for Moderate and Large Size Samples, *Biometrika* 58:341-348.

D'Agostino, R.B. and M.A. Stephens. 1986. *Goodness-of-Fit Techniques*, Marcel Dekker, New York.

EPA 1992a. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities – Addendum to Interim Final Guidance*. U.S. Environmental Protection Agency, Office of Solid Waste, Permits and State Programs Division, Washington, DC.

EPA 1992b. *User Documentation: GRITS/STAT, v4.2. EPA/625/11-91/002*, U.S. Environmental Protection Agency, Center for Environmental Research Information, Cincinnati, OH.

EPA 1993. *Data Quality Objectives Process for Superfund*, Interim Final Guidance, EPA540-R-93-071, U.S. Environmental Protection Agency, Office of Solid Waste and Emergency Response, Washington, DC.

EPA 1994a. *Guidance for the Data Quality Objectives Process*, EPA QA/G-4, EPA/600/R-96/055, U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

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

EPA 1996. *Guidance for Data Quality Assessment, Practical Methods for Data Analysis, EPA QA/G-9, QA96 Version*, EPA/600/R-96/084, U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

EPA 1997. *Data Quality Evaluation Statistical Toolbox (DataQUEST) User's Guide*, EPA QA/G-9D, QA96 Version, EPA/600/R-96/085, U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

EPA 1998. *Guidance for Data Quality Assessment, Practical Methods for Data analysis, EPA QA/G-9, QA97 Update*, EPA/600/R-96/084. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

EPA 1999. *Guidance on Sampling Design to Support Quality Assurance Project Plans*, EPA QA/G-5S, U.S. Environmental Protection Agency, Washington, DC. (In preparation).

Filliben, J.J. 1975. The Probability Plot Correlation Coefficient Test for Normality, *Technometrics* 17:111-117.

Gehan, E.A. 1965. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika* 52:203-223.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*, Van Nostrand Reinhold, New York (now published by Wiley & Sons, New York, 1997).

Gilbert, R.O. and J.C. Simpson. 1990. Statistical Sampling and Analysis Issues and Needs for Testing Attainment of Background-Based Cleanup Standards at Superfund Sites, In *Proceedings of the Workshop on Superfund Hazardous Waste: Statistical Issues in Characterizing a Site: Protocols, Tools, and Research Needs* (H. Lacayo, R.J. Nadeau, G.P. Patil, and L Zaragoza, editors), pp. 1-16, Center for Statistical Ecology and Environmental Statistics, The Pennsylvania State University, University Park, PA.

Helsel, D.R. and R.M. Hirsch. 1992. *Statistical Methods in Water Resources*, Elsevier, New York, NY.

Iman, R.L. and W.J. Conover. 1983. *A Modern Approach to Statistics*, Wiley & Sons, New York.

MARSSIM 1997. *Multi-Agency Radiation Survey and Site Investigation Manual (MARSSIM)*, NUREG-1575, EPA 402-R-97-016, Nuclear Regulatory Commission and the U.S. Environmental Protection Agency, Washington DC.

Michael, J.R. and W.R. Schucany. 1986. Analysis of Data from Censored Samples, pp. 461-496, In: *Goodness-of-Fit Techniques*, (R.B. D'Agostino and M.A. Stephens, Editors), Marcel Dekker, New York.

Millard, S.P. 1997. *EnvironmentalStat for S-Plus, User's Manual*, Version 1.0, Probability, Statistics & Information, 7723 44th Ave. N.E., Seattle, WA 98115-5117.

Navy. 1998. *Procedural Guidance for Statistically Analyzing Environmental Background Data*, Department of the Navy, Southwest Division, Naval Facilities Engineering Command, San Diego, CA.

O'Brien, R.F. and R.O. Gilbert. 1997. Comparing Sites to Background, *Environmental Testing & Analysis*, September/October issue, pp. 10-13.

Palachek, A.D., D.R. Weier, T.R. Gatliffe, D.M. Splett, and D.K. Sullivan. 1993. *Statistical Methodology for Determining Contaminants of Concern by Comparison of Background and Site Data with Applications to Operable Unit 2*, SA-93-010, Internal Report, Statistical Applications, EG&G Rocky Flats Inc., Rocky Flats Plant, Golden, CO.

7.0 TABLES

Table A.1 Cumulative Standard Normal Distribution (Values of the Probability f Corresponding to the Value Z_f of a Standard Normal Random Variable)

Z_f	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5674	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Table A.2 Critical Values for the Extreme Value Test For Outliers (Dixon Test)

	Level of Significance a		
n	**0.10**	**0.05**	**0.01**
3	0.886	0.941	0.988
4	0.679	0.765	0.889
5	0.557	0.642	0.780
6	0.482	0.560	0.698
7	0.434	0.507	0.637
8	0.479	0.554	0.683
9	0.441	0.512	0.635
10	0.409	0.477	0.597
11	0.517	0.576	0.679
12	0.490	0.546	0.642
13	0.467	0.521	0.615
14	0.492	0.546	0.641
15	0.472	0.525	0.616
16	0.454	0.507	0.595
17	0.438	0.490	0.577
18	0.424	0.475	0.561
19	0.412	0.462	0.547
20	0.401	0.450	0.535
21	0.391	0.440	0.524
22	0.382	0.430	0.514
23	0.374	0.421	0.505
24	0.367	0.413	0.497
25	0.360	0.406	0.489

Table A.3 Critical Values for the Discordance Test for Outliers

n	Level of Significance 0.01	0.05
3	1.155	1.153
4	1.492	1.463
5	1.749	1.672
6	1.944	1.822
7	2.097	1.938
8	2.221	2.032
9	2.323	2.110
10	2.410	2.176
11	2.485	2.234
12	2.550	2.285
13	2.607	2.331
14	2.659	2.371
15	2.705	2.409
16	2.747	2.443
17	2.785	2.475
18	2.821	2.504
19	2.854	2.532
20	2.884	2.557
21	2.912	2.580
22	2.939	2.603
23	2.963	2.624
24	2.987	2.644
25	3.009	2.663
26	3.029	2.681
27	3.049	2.698
28	3.068	2.714
29	3.085	2.730
30	3.103	2.745
31	3.119	2.759
32	3.135	2.773

n	Level of Significance 0.01	0.05
33	3.150	2.786
34	3.164	2.799
35	3.178	2.811
36	3.191	2.823
37	3.204	2.835
38	3.216	2.846
39	3.228	2.857
40	3.240	2.866
41	3.251	2.877
42	3.261	2.887
43	3.271	2.896
44	3.282	2.905
45	3.292	2.914
46	3.302	2.923
47	3.310	2.931
48	3.319	2.940
49	3.329	2.948
50	3.336	2.956

Table A.4 Approximate Critical Values for the Rosner Test for Outliers

n	r	a 0.05	a 0.01
25	1	2.82	3.14
	2	2.80	3.11
	3	2.78	3.09
	4	2.76	3.06
	5	2.73	3.03
	10	2.59	2.85
26	1	2.84	3.16
	2	2.82	3.14
	3	2.80	3.11
	4	2.78	3.09
	5	2.76	3.06
	10	2.62	2.89
27	1	2.86	3.18
	2	2.84	3.16
	3	2.82	3.14
	4	2.80	3.11
	5	2.78	3.09
	10	2.65	2.93
28	1	2.88	3.20
	2	2.86	3.18
	3	2.84	3.16
	4	2.82	3.14
	5	2.80	3.11
	10	2.68	2.97
29	1	2.89	3.22
	2	2.88	3.20
	3	2.86	3.18
	4	2.84	3.16
	5	2.82	3.14
	10	2.71	3.00
30	1	2.91	3.24
	2	2.89	3.22
	3	2.88	3.20
	4	2.86	3.18
	5	2.84	3.16
	10	2.73	3.03
31	1	2.92	3.25
	2	2.91	3.24
	3	2.89	3.22
	4	2.88	3.20
	5	2.86	3.18
	10	2.76	3.06
46	1	3.09	3.45
	2	3.09	3.44
	3	3.08	3.43
	4	3.07	3.41
	5	3.06	3.40
	10	3.00	3.34

n	r	a 0.05	a 0.01
32	1	2.94	3.27
	2	2.92	3.25
	3	2.91	3.24
	4	2.89	3.22
	5	2.88	3.20
	10	2.78	3.09
33	1	2.95	3.29
	2	2.94	3.27
	3	2.92	3.25
	4	2.91	3.24
	5	2.89	3.22
	10	2.80	3.11
34	1	2.97	3.30
	2	2.95	3.29
	3	2.94	3.27
	4	2.92	3.25
	5	2.91	3.24
	10	2.82	3.14
35	1	2.98	3.32
	2	2.97	3.30
	3	2.95	3.29
	4	2.94	3.27
	5	2.92	3.25
	10	2.84	3.16
36	1	2.99	3.33
	2	2.98	3.32
	3	2.97	3.30
	4	2.95	3.29
	5	2.94	3.27
	10	2.86	3.18
37	1	3.00	3.34
	2	2.99	3.33
	3	2.98	3.32
	4	2.97	3.30
	5	2.95	3.29
	10	2.88	3.20
38	1	3.01	3.36
	2	3.00	3.34
	3	2.99	3.33
	4	2.98	3.32
	5	2.97	3.30
	10	2.91	3.22
70	1	3.26	3.62
	2	3.25	3.62
	3	3.25	3.61
	4	3.24	3.60
	5	3.24	3.60
	10	3.21	3.57

n	r	a 0.05	a 0.01
39	1	3.03	3.37
	2	3.01	3.36
	3	3.00	3.34
	4	2.99	3.33
	5	2.98	3.32
	10	2.91	3.24
40	1	3.04	3.38
	2	3.03	3.37
	3	3.01	3.36
	4	3.00	3.34
	5	2.99	3.33
	10	2.92	3.25
41	1	3.05	3.39
	2	3.04	3.38
	3	3.03	3.37
	4	3.01	3.36
	5	3.00	3.34
	10	2.94	3.27
42	1	3.06	3.40
	2	3.05	3.39
	3	3.04	3.38
	4	3.03	3.37
	5	3.01	3.36
	10	2.95	3.29
43	1	3.07	3.41
	2	3.06	3.40
	3	3.05	3.39
	4	3.04	3.38
	5	3.03	3.37
	10	2.97	3.30
44	1	3.08	3.43
	2	3.07	3.41
	3	3.06	3.40
	4	3.05	3.39
	5	3.04	3.38
	10	2.98	3.32
45	1	3.09	3.44
	2	3.08	3.43
	3	3.07	3.41
	4	3.06	3.40
	5	3.05	3.39
	10	2.99	3.33
250	1	3.67	4.04
	5	3.67	4.04
	10	3.66	4.03

		a	
n	**r**	**0.05**	**0.01**
47	1	3.10	3.46
	2	3.09	3.45
	3	3.09	3.44
	4	3.08	3.43
	5	3.07	3.41
	10	3.01	3.36
48	1	3.11	3.46
	2	3.10	3.46
	3	3.09	3.45
	4	3.09	3.44
	5	3.08	3.43
	10	3.03	3.37
49	1	3.12	3.47
	2	3.11	3.46
	3	3.10	3.46
	4	3.09	3.45
	5	3.09	3.44
	10	3.04	3.38
50	1	3.13	3.48
	2	3.12	3.47
	3	3.11	3.46
	4	3.10	3.46
	5	3.09	3.45
	10	3.05	3.39
60	1	3.20	3.56
	2	3.19	3.55
	3	3.19	3.55
	4	3.18	3.54
	5	3.17	3.53
	10	3.14	3.49

		a	
n	**r**	**0.05**	**0.01**
80	1	3.31	3.67
	2	3.30	3.67
	3	3.30	3.66
	4	3.29	3.66
	5	3.29	3.65
	10	3.26	3.63
90	1	3.35	3.72
	2	3.34	3.71
	3	3.34	3.71
	4	3.34	3.70
	5	3.33	3.70
	10	3.31	3.68
100	1	3.38	3.75
	2	3.38	3.75
	3	3.38	3.75
	4	3.37	3.74
	5	3.37	3.74
	10	3.35	3.72
150	1	3.52	3.89
	2	3.51	3.89
	3	3.51	3.89
	4	3.51	3.88
	5	3.51	3.88
	10	3.50	3.87
200	1	3.61	3.98
	2	3.60	3.98
	3	3.60	3.97
	4	3.60	3.97
	5	3.60	3.97
	10	3.59	3.96

		a	
n	**r**	**0.05**	**0.01**
300	1	3.72	4.09
	5	3.72	4.09
	10	3.71	4.09
350	1	3.77	4.14
	5	3.76	4.13
	10	3.76	4.13
400	1	3.80	4.17
	5	3.80	4.17
	10	3.80	4.16
450	1	3.84	4.20
	5	3.83	4.20
	10	3.83	4.20
500	1	3.86	4.23
	5	3.86	4.23
	10	3.86	4.22

Table A.5 Values of the Parameter l for the Cohen Estimates of the Mean and Variance of Normally Distributed Data Sets That Contain Non-Detects

	h											
g	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**	**.10**	**.15**	**.20**
00	.010100	.020400	.030902	.041583	.052507	.063625	.074953	.08649	.09824	.11020	.17342	.24268
05	.010551	.021294	.032225	.043350	.054670	.066159	.077909	.08983	.10197	.11431	.17925	.25033
10	.010950	.022082	.033398	.044902	.056596	.068483	.080563	.09285	.10534	.11804	.18479	.25741
15	.011310	.022798	.034466	.046318	.058356	.070586	.083009	.09563	.10845	.12148	.18985	.26405
20	.011642	.023459	.035453	.047829	.059990	.072539	.085280	.09822	.11135	.12469	.19460	.27031
25	.011952	.024076	.036377	.048858	.061522	.074372	.087413	.10065	.11408	.12772	.19910	.2762
30	.012243	.024658	.037249	.050018	.062969	.076106	.089433	.10295	.11667	.13059	.20338	.2819
35	.012520	.025211	.038077	.051120	.064345	.077736	.091355	.10515	.11914	.13333	.20747	.2873
40	.012784	.025738	.038866	.052173	.065660	.079332	.093193	.10725	.12150	.13595	.21129	.2925
45	.013036	.026243	.039624	.053182	.066921	.080845	.094958	.10926	.12377	.13847	.21517	.2976
50	.013279	.026728	.040352	.054153	.068135	.082301	.096657	.11121	.12595	.14090	.21882	.3025
55	.013513	.027196	.041054	.055089	.069306	.083708	.098298	.11208	.12806	.14325	.22225	.3072
60	.013739	.027849	.041733	.055995	.070439	.085068	.099887	.11490	.13011	.14552	.22578	.3118
65	.013958	.028087	.042391	.056874	.071538	.086388	.10143	.11666	.13209	.14773	.22910	.3163
70	.014171	.028513	.043030	.057726	.072505	.087670	.10292	.11837	.13402	.14987	.23234	.3206
75	.014378	.029927	.043652	.058556	.073643	.088917	.10438	.12004	.13590	.15196	.23550	.32489
80	.014579	.029330	.044258	.059364	.074655	.090133	.10580	.12167	.13775	.15400	.23858	.32903
85	.014773	.029723	.044848	.060153	.075642	.091319	.10719	.12225	.13952	.15599	.24158	.33307
90	.014967	.030107	.045425	.060923	.075606	.092477	.10854	.12480	.14126	.15793	.24452	.33703
95	.015154	.030483	.045989	.061676	.077549	.093611	.10987	.12632	.14297	.15983	.24740	.34091
1.00	.01533 8	.03085 0	.04654 0	.06241 3	.07847 1	.09472 0	.11116	.1278 0	.1446 5	.1617 0	.2502 2	.3447 1

	h											
g	**.25**	**.30**	**.35**	**.40**	**.45**	**.50**	**.55**	**.60**	**.65**	**.70**	**.80**	**.90**
.00	.31862	.4021	.4941	.5961	.7096	.8388	.9808	1.145	1.336	1.561	2.176	3.283
.05	.32793	.4130	.5066	.6101	.7252	.8540	.9994	1.166	1.358	1.585	2.203	3.314
.10	.33662	.4233	.5184	.6234	.7400	.8703	1.017	1.185	1.379	1.608	2.229	3.345
.15	.34480	.4330	.5296	.6361	.7542	.8860	1.035	1.204	1.400	1.630	2.255	3.376
.20	.35255	.4422	.5403	.6483	.7673	.9012	1.051	1.222	1.419	1.651	2.280	3.405
.25	.35993	.4510	.5506	.6600	.7810	.9158	1.067	1.240	1.439	1.672	2.305	3.435
.30	.36700	.4595	.5604	.6713	.7937	.9300	1.083	1.257	1.457	1.693	2.329	3.464
.35	.37379	.4676	.5699	.6821	.8060	.9437	1.098	1.274	1.475	1.713	2.353	3.492
.40	.38033	.4735	.5791	.6927	.8179	.9570	1.113	1.290	1.494	1.732	2.376	3.520
.45	.38665	.4831	.5880	.7029	.8295	.9700	1.127	1.306	1.511	1.751	2.399	3.547
.50	.39276	.4904	.5967	.7129	.8408	.9826	1.141	1.321	1.528	1.770	2.421	3.575
.55	.39679	.4976	.6061	.7225	.8517	.9950	1.155	1.337	1.545	1.788	2.443	3.601
.60	.40447	.5045	.6133	.7320	.8625	1.007	1.169	1.351	1.561	1.806	2.465	3.628
.65	.41008	.5114	.6213	.7412	.8729	1.019	1.182	1.368	1.577	1.824	2.486	3.654
.70	.41555	.5180	.6291	.7502	.8832	1.030	1.195	1.380	1.593	1.841	2.507	3.679
.75	.42090	.5245	.6367	.7590	.8932	1.042	1.207	1.394	1.608	1.851	2.528	3.705
.80	.42612	.5308	.6441	.7676	.9031	1.053	1.220	1.408	1.624	1.875	2.548	3.730
.85	.43122	.5370	.6515	.7781	.9127	1.064	1.232	1.422	1.639	1.892	2.568	3.754
.90	.43622	.5430	.6586	.7844	.9222	1.074	1.244	1.435	1.653	1.908	2.588	3.779
.95	.44112	.5490	.6656	.7925	.9314	1.085	1.255	1.448	1.668	1.924	2.607	3.803
1.00	.44592	.5548	.6724	.8005	.9406	1.095	1.287	1.461	1.882	1.940	2.626	3.827

Table A.6 Coefficients a_k for the Shapiro-Wilk W Test for Normality

k\n	2	3	4	5	6	7	8	9	10
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5868	0.5739
2	-	0.0000	0.1677	0.2413	0.28D6	0.3031	0.3164	0.3244	0.3291
3	-	-	-	0.0000	0.0875	0.1401	0.1743	0.1976	0.2141
4	-	-	-	-	-	0.0000	0.0561	0.0947	0.1224
5	-	-	-	-	-	-	-	0.0000	0.0399

k\n	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0,2540	0.2553	0.2561	0.2565
4	0.1429	0.1506	0.1707	0.1802	0.1876	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0 0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7	-	-	0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8	-	-	-	-	0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9	-	-	-	-	-	-		0.0163	0.0303	0.0422
10	-	-	-	-	-	-	-	-	0.0000	0.0140

k\n	21	22	23	24	25	26	27	28	29	30
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254
2	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043	0.3018	0.2992	0.2968	0.2944
3	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533	0.2522	0.2510	0.2499	0.2487
4	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151	0.2152	0.2151	0.2150	0.2148
5	0.1736	0.1764	0.1787	0.1007	0.1822	0.1836	0.1840	0.1857	0.1864	0.1870
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630
7	0.1092	0.1150	0.1201	0.1245	0.1263	0.1316	0.1346	0.1372	0.1395	0.1415
8	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089	0.1128	0.1162	0.1192	0.1219
9	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876	0.0923	0.0965	0.1002	0.1036
10	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672	0.0728	0.0778	0.0822	0.0862
11	0.0000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697
12	-	-	0.0000	0.0107	0.0200	0.0284	0.0358	0.0424	0.0483	0.0537
13	-	-	-	-	0.0000	0.0094	0.0178	0.0253	0.0320	0.0381
14	-	-	-	-	-	-	0.0000	0.0084	0.0159	0.0227
15	-	-	-	-	-	-	-	-	0.0000	0.0076

Table A.7. Quantiles of the Shapiro-Wilk W Test for Normality

n	$W_{0.01}$	$W_{0.02}$	$W_{0.05}$	$W_{0.10}$	$W_{0.50}$
3	0.753	0.756	0.767	0.789	0.859
4	0.687	0.707	0.748	0.792	0.935
5	0.686	0.715	0.762	0.806	0.927
6	0.713	0.743	0.788	0.826	0.927
7	0.730	0.760	0.803	0.838	0.928
8	0.749	0.778	0.818	0.851	0.932
9	0.764	0.791	0.829	0.859	0.935
10	0.781	0.806	0.842	0.869	0.938
11	0.792	0.817	0.850	0.876	0.940
12	0.805	0.828	0.859	0.883	0.943
13	0.814	0.837	0.866	0.889	0.945
14	0.825	0.846	0.874	0.895	0.947
15	0.835	0.855	0.881	0.901	0.950
16	0.844	0.863	0.887	0.906	0.952
17	0.851	0.869	0.892	0.910	0.954
18	0.858	0.874	0.897	0.914	0.956
19	0.863	0.879	0.901	0.917	0.957
20	0.868	0.886	0.905	0.920	0.969
21	0.873	0.884	0.908	0.923	0.960
22	0.878	0.892	0.911	0.926	0.961
23	0.881	0.895	0.914	0.928	0.962
24	0.884	0.898	0.916	0.930	0.963
25	0.886	0.901	0.918	0.931	0.964
26	0.891	0.904	0.920	0.933	0.965
27	0.894	0.906	0.923	0.935	0.965
28	0.896	0.908	0.924	0.936	0.966
29	0.898	0.910	0.926	0.937	0.966
30	0.900	0.912	0.927	0.939	0.967
31	0.902	0.914	0.929	0.940	0.967
32	0.904	0.915	0.930	0.941	0.968
33	0.906	0.917	0.931	0.942	0.968
34	0.908	0.919	0.933	0..943	0.969
35	0.910	0.920	0.934	0.944	0.969
36	0.912	0.922	0.935	0.945	0.970
37	0.914	0.924	0.936	0.946	0.970
38	0.916	0.925	0.938	0.947	0.971
39	0.917	0.927	0.939	0.948	0.971
40	0.919	0.928	0.940	0.949	0.972
41	0.920	0.929	0.941	0.950	0.972
42	0.922	0.930	0.942	0.951	0.972
43	0.923	0.932	0.943	0.951	0.973
44	0.924	0.933	0.944	0.952	0.973
45	0.926	0.934	0.945	0.953	0.973
46	0.927	0.935	0.945	0.953	0.974
47	0.928	0.936	0.946	0.954	0.974
48	0.929	0.937	0.947	0.954	0.974
49	0.929	0.937	0.947	0.955	0.974
50	0.930	0.938	0.947	0.955	0.974

Table A.8 Quantiles of the D'Agostino Test for Normality (Values of Y such that 100p% of the Distribution of Y is Less than Y_p)

n	$Y_{0.005}$	$Y_{0.01}$	$Y_{0.025}$	$Y_{0.05}$	$Y_{0.10}$	$Y_{0.90}$	$Y_{0.95}$	$Y_{0.975}$	$Y_{0.99}$	$Y_{0.995}$
50	-3.949	-3.442	-2.757	-2.220	-1.661	0.759	0.923	1.038	1.140	1.192
60	-3.846	-3.360	-2.699	-2.179	-1.634	0.807	0.986	1.115	1.236	1.301
70	-3.762	-3.293	-2.652	-2.146	-1.612	0.844	1.036	1.176	1.312	1.388
80	-3.693	-3.237	-2.613	-2.118	-1.594	0.874	1.076	1.226	1.374	1.459
90	-3.635	-3.100	-2.580	-2.095	-1.579	0.899	1.109	1.268	1.426	1.518
100	-3.584	-3.150	-2.552	-2.075	-1.566	0.920	1.137	1.303	1.470	1.569
150	-3.409	-3.009	-2.452	-2.004	-1.520	0.990	1.233	1.423	1.623	1.746
200	-3.302	-2.922	-2.391	-1.960	-1.491	1.032	1.290	1.496	1.715	1.853
250	-3.227	-2.861	-2.348	-1.926	-1.471	1.060	1.328	1.545	1.779	1.927
300	-3.172	-2.816	-2.316	-1.906	-1.456	1.080	1.357	1.528	1.826	1.983
350	-3.129	-2.781	-2.291	-1.888	-1.444	1.096	1.379	1.610	1.863	2.026
400	-3.094	-2.753	-2.270	-1.873	-1.434	1.108	1.396	1.633	1.893	2.061
450	-3.064	-2.729	-2.253	-1.861	-1.426	1.119	1.411	1.652	1.918	2.090
500	-3.040	-2.709	-2.239	-1.850	-1.419	1.127	1.423	1.668	1.938	2.114
550	-3.019	-2.691	-2.226	-1.841	-1.413	1.135	1.434	1.682	1.957	2.136
600	-3.000	-2.676	-2.215	-1.833	-1.408	1.141	1.443	1.694	1.972	2.154
650	-2.984	-2.663	-2.206	-1.826	-1.403	1.147	1.451	1.704	1.986	2.171
700	-2.969	-2.651	-2.197	-1.820	-1.399	1.152	1.458	1.714	1.999	2.185
750	-2.956	-2.640	-2.189	-1.814	-1.395	1.157	1.465	1.722	2.010	2.199
800	-2.944	-2.630	-2.182	-1.809	-1.392	1.161	1.471	1.730	2.020	2.211
850	-2.933	-2.621	-2.176	-1.804	-1.389	1.165	1.476	1.737	2.029	2.221
900	-2.923	-2.613	-2.170	-1.800	-1.386	1.168	1.481	1.743	2.037	2.231
950	-2.914	-2.605	-2.164	-1.796	-1.383	1.171	1.485	1.749	2.045	2.241
1000	-2.906	-2.599	-2.159	-1.792	-1.381	1.174	1.489	1.754	2.052	2.249

Table A.9 Critical Values for the Slippage Test for a = 0.01

Number of Site Measurements, n

Number of Background Measeurments, m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
2	/	/	/	/	/	/	/	/	/	/	/	/	13	14	15	16	17	18	19	20	21	22	23	23	24
3	/	/	/	/	/	/	7	8	9	10	11	11	12	13	14	15	15	16	17	18	18	19	20	21	22
4	/	/	/	/	5	6	7	8	8	9	10	10	11	12	12	13	14	14	15	16	16	17	18	19	19
5	/	/	/	4	5	6	6	7	8	8	9	9	10	11	11	12	12	13	14	14	15	15	16	17	17
6	/	/	/	4	5	5	6	6	7	8	8	9	9	10	10	11	11	12	12	13	14	14	15	15	16
7	/	/	3	4	5	5	6	6	7	7	8	8	9	9	10	10	10	11	11	12	12	13	13	14	14
8	/	/	3	4	4	5	5	6	6	7	7	8	8	8	9	9	10	10	11	11	12	12	12	13	13
9	/	/	3	4	4	5	5	5	6	6	7	7	8	8	8	9	9	10	10	10	11	11	12	12	12
10	/	/	3	4	4	4	5	5	6	6	6	7	7	7	8	8	9	9	9	10	10	11	11	11	12
11	/	/	3	3	4	4	5	5	5	6	6	6	7	7	7	8	8	9	9	9	10	10	10	11	11
12	/	/	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	9	10	10	10
13	/	2	3	3	4	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8	9	9	9	10	10
14	/	2	3	3	4	4	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8	9	9	9	9
15	/	2	3	3	3	4	4	4	5	5	5	5	6	6	6	7	7	7	7	8	8	8	9	9	9
16	/	2	3	3	3	4	4	4	4	5	5	5	6	6	6	6	7	7	7	7	8	8	8	8	9
17	/	2	3	3	3	4	4	4	4	5	6	6	6	6	6	6	6	7	7	7	7	8	8	8	8
18	/	2	3	3	3	3	4	4	4	5	5	5	5	5	6	6	6	6	7	7	7	7	8	8	8
19	/	2	3	3	3	3	4	4	4	4	5	5	5	6	6	6	6	6	6	7	7	7	7	8	8
20	/	2	3	3	3	3	4	4	4	4	5	5	5	5	5	6	6	6	6	7	7	7	7	7	8
21	/	2	3	3	3	3	4	4	4	4	4	5	5	5	5	6	6	6	6	6	7	7	7	7	7
22	/	2	3	3	3	3	3	4	4	4	4	5	5	5	5	5	6	6	6	6	6	7	7	7	7
23	/	2	2	3	3	3	3	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	7	7	7
24	/	2	2	3	3	3	3	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	6	7	7
25	/	2	2	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	6	7
26	/	2	2	3	3	3	3	3	4	4	4	4	4	5	5	5	5	5	5	6	6	6	6	6	6
27	/	2	2	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	6
28	/	2	2	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	6	6	6	6	6
29	/	2	2	3	3	3	3	3	3	4	4	4	4	4	5	5	5	5	5	5	5	6	6	6	6
30	/	2	2	3	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	6	6	6	6
31	/	2	2	3	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	5	6	6	6
32	/	2	2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	5	5	6	6
33	/	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	6	6	6
34	/	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	5	6	6
35	/	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5
36	/	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5
37	/	2	2	2	3	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	5	5
38	/	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5	5
39	/	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5	5
40	/	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5
41	/	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5
42	/	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5
43	/	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	5	5	5	5
44	/	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5
45	/	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	5	5	5
46	/	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	5	5	5
47	/	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	5	5	5
48	/	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	5	5
49	/	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	5
50	/	2	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	5

Table A.9 Critical Values for the Slippage Test for a = 0.01 (continued)

Number of Site Measurments, n

Number of Background Measurements, m	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
2	25	26	27	28	29	30	31	32	32	33	34	35	36	37	38	39	40	41	41	42	43	44	45	46	47
3	22	23	24	25	26	26	27	28	29	30	30	31	32	33	33	34	35	36	37	37	38	39	40	40	41
4	20	21	21	22	23	23	24	25	25	26	27	27	28	29	30	30	31	32	32	33	34	34	35	35	36
5	18	18	19	20	20	21	22	22	23	23	24	25	25	26	26	27	28	28	29	29	30	31	31	32	32
6	16	17	17	18	18	19	19	20	21	21	22	22	23	23	24	24	25	25	26	26	27	28	28	29	29
7	15	15	16	16	17	17	18	18	19	19	20	20	21	21	22	22	23	23	24	24	25	25	26	26	26
8	14	14	15	15	16	16	16	17	17	18	18	19	19	19	20	20	21	21	22	22	23	23	23	24	24
9	13	13	14	14	14	15	15	16	16	16	17	17	18	18	18	19	19	20	20	20	21	21	22	22	22
10	12	12	13	13	13	14	14	15	15	15	16	16	16	17	17	18	18	18	19	19	19	20	20	21	21
11	11	12	12	12	13	13	13	14	14	14	15	15	15	16	16	16	17	17	18	18	18	19	19	19	20
12	11	11	11	12	12	12	13	13	13	14	14	14	15	15	15	16	16	16	17	17	17	17	18	18	18
13	10	10	11	11	11	12	12	12	13	13	13	14	14	14	14	15	15	15	16	16	16	17	17	17	17
14	10	10	10	11	11	11	11	12	12	12	13	13	13	13	14	14	14	15	15	15	15	16	16	16	17
15	9	10	10	10	10	11	11	11	11	12	12	12	13	13	13	13	14	14	14	14	15	15	15	15	16
16	9	9	9	10	10	10	10	11	11	11	11	12	12	12	13	13	13	13	14	14	14	14	15	15	15
17	9	9	9	9	10	10	10	10	11	11	11	11	12	12	12	12	12	13	13	13	13	14	14	14	14
18	8	9	9	9	9	9	10	10	10	10	11	11	11	11	12	12	12	12	12	13	13	13	13	14	14
19	8	8	8	9	9	9	9	10	10	10	10	10	11	11	11	11	12	12	12	12	13	13	13	13	13
20	8	8	8	8	9	9	9	9	9	10	10	10	10	11	11	11	11	11	12	12	12	12	12	13	13
21	8	8	8	8	8	9	9	9	9	9	10	10	10	10	10	11	11	11	11	11	12	12	12	12	12
22	7	8	8	8	8	8	9	9	9	9	9	9	10	10	10	10	10	11	11	11	11	11	12	12	12
23	7	7	8	8	8	8	8	8	9	9	9	9	9	10	10	10	10	10	11	11	11	11	11	11	12
24	7	7	7	8	8	8	8	8	8	9	9	9	9	9	9	10	10	10	10	10	11	11	11	11	11
25	7	7	7	7	8	8	8	8	8	8	9	9	9	9	9	9	10	10	10	10	10	10	11	11	11
26	7	7	7	7	7	7	8	8	8	8	8	8	9	9	9	9	9	9	10	10	10	10	10	10	11
27	7	7	7	7	7	7	7	8	8	8	8	8	8	9	9	9	9	9	9	10	10	10	10	10	10
28	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9	9	9	9	10	10	10	10
29	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	9	9	9	9	9	9	9	10	10	10
30	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9	9	9	9	9	10
31	6	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9	9	9	9	9
32	6	6	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9	9	9	9
33	6	6	6	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9	9	9
34	6	6	6	6	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	8	9	9	9
35	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	8	8	8	8	8	8	8	8	9	9
36	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	8	8	8	8	8	8	8	8	9
37	5	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	8	8	8	8	8	8	8	8
38	5	5	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	8	8	8	8	8	8	8
39	5	5	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	8	8	8	8	8	8
40	5	5	5	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	8	8	8	8
41	5	5	5	5	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	8	8	8
42	5	5	5	5	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	8	8
43	5	5	5	5	5	5	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	8
44	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7
45	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7
46	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7
47	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7
48	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7
49	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7
50	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7

Table A.10 Critical Values for the Slippage Test for $\alpha = 0.05$

Number of Site Measurements, n

Number of Background Measurements, m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	20	21	22	23	24	25
2	/	/	/	/	5	6	7	8	9	9	10	11	12	13	13	14	15	16	16	17	18	19	20	20	21
3	/	/	/	4	5	5	6	7	7	8	9	9	10	11	11	12	12	13	14	14	15	16	16	17	18
4	/	/	3	4	4	5	5	6	6	7	8	8	9	9	10	10	11	11	12	12	13	13	14	14	15
5	/	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	9	10	10	11	11	12	12	13	13
6	/	2	3	3	4	4	4	5	5	6	6	6	7	7	8	8	8	9	9	10	10	10	11	11	12
7	/	2	3	3	3	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9	9	10	10	11
8	/	2	3	3	3	4	4	4	5	5	5	6	6	6	6	7	7	7	8	8	8	9	9	9	10
9	/	2	2	3	3	3	4	4	4	5	5	5	5	6	6	6	7	7	7	7	8	8	8	9	9
10	/	2	3	3	3	3	4	4	4	4	5	5	5	5	6	6	6	6	7	7	7	8	8	8	8
11	/	2	2	3	3	3	3	4	4	4	4	5	5	5	5	6	6	6	6	7	7	7	7	8	8
12	/	2	2	3	3	3	3	3	4	4	4	4	5	5	5	5	6	6	6	6	6	7	7	7	7
13	/	2	2	2	3	3	3	3	4	4	4	4	4	5	5	5	5	6	6	6	6	6	7	7	7
14	/	2	2	2	3	3	3	3	3	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	7
15	/	2	2	2	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6
16	/	2	2	2	2	3	3	3	3	3	4	4	4	4	4	5	5	5	5	5	5	6	6	6	6
17	/	2	2	2	2	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	6	6	6
18	/	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	5	6
19	/	2	2	2	2	3	3	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	5
20	1	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	5
21	1	2	2	2	2	2	3	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5
22	1	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5
23	1	2	2	2	2	2	2	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5
24	1	2	2	2	2	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5
25	1	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5
26	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5
27	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4
28	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
29	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
30	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4
31	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4
32	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4
33	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4
34	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4
35	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4
36	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4
37	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4
38	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4
39	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	4	4
40	1	1	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	4
41	1	1	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	4
42	1	1	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3
43	1	1	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3
44	1	1	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3
45	1	1	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3
46	1	1	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3
47	1	1	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3
48	1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3
49	1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3
50	1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3

Table A.10 Critical Values for the Slippage Test for a = 0.05 (continued)

Number of Site Measurements, n

SITE	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
1	26	27	28	29	30	31	32	33	34	35	36	37	38	39	39	40	41	42	43	44	45	46	47	48	49
2	22	23	23	24	25	26	26	27	28	29	30	30	31	32	33	33	34	35	36	37	37	38	39	40	40
3	18	19	19	20	21	21	22	23	23	24	24	25	26	26	27	28	28	29	30	30	31	31	32	33	33
4	15	16	17	17	18	18	19	19	20	20	21	21	22	22	23	23	24	24	25	26	26	27	27	28	28
5	14	14	14	15	15	16	16	17	17	18	18	18	19	19	20	20	21	21	22	22	23	23	23	24	24
6	12	12	13	13	14	14	14	15	15	16	16	16	17	17	18	18	18	19	19	20	20	20	21	21	21
7	11	11	12	12	12	13	13	13	14	14	14	15	15	15	16	16	16	17	17	18	18	18	19	19	19
8	10	10	11	11	11	12	12	12	13	13	13	13	14	14	14	15	15	15	16	16	16	17	17	17	17
9	9	9	10	10	10	11	11	11	11	12	12	12	13	13	13	13	14	14	14	15	15	16	16	16	16
10	9	9	9	9	10	10	10	10	11	11	11	11	12	12	12	12	13	13	13	14	14	14	14	15	15
11	8	8	9	9	9	9	9	10	10	10	10	11	11	11	11	12	12	12	12	13	13	13	13	14	14
12	8	8	8	8	8	9	9	9	9	10	10	10	10	10	11	11	11	11	12	12	12	12	12	13	13
13	7	7	8	8	8	8	8	9	9	9	9	9	10	10	10	10	11	11	11	11	11	12	12	12	12
14	7	7	7	7	8	8	8	8	8	9	9	9	9	9	10	10	10	10	10	11	11	11	11	11	12
15	7	7	7	7	7	7	8	8	8	8	9	9	9	9	9	9	9	10	10	10	10	10	11	11	11
16	6	6	7	7	7	7	7	7	8	8	8	8	8	9	9	9	9	9	9	10	10	10	10	10	10
17	6	6	6	7	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9	9	9	9	10	10	10
18	6	6	6	6	6	7	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9	9	9	9	10
19	6	6	6	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9	9	9
20	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	9	9	9
21	5	5	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	9
22	5	5	5	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	8	8	8	8	8	8
23	5	5	5	5	5	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	8	8	8	8
24	5	5	5	5	5	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	8	8
25	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	8
26	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7
27	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7
28	4	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	7	7	7	7	7
29	4	4	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	7	7	7
30	4	4	4	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	7	7
31	4	4	4	4	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6	6
32	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6
33	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6
34	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6
35	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6
36	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6
37	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	6	6	6
38	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	6	6
39	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	6
40	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5
41	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5
42	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5
43	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5
44	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5
45	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5
46	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5
47	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5
48	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5
49	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5
50	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5

Number of Background Measurements, m

Table A.11 Values of r, k and a for the Quantile Test when a is Approximately Equal to 0.01

Number of Site Measurements, n

Number of Reference (Background) Measurements, m

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
5	**r, k** **a**		11,1 0.008	13,13 0.015	16,16 0.014	19,19 0.013	22,22 0.013	25,25 0.013	28,28 0.012											
10		6,6 0.005	7,7 0.013	9,9 0.012	11,11 0.011	13,13 0.010	14,14 0.014	16,16 0.013	18,18 0.012	19,19 0.015	21,21 0.014	23,23 0.013	25,25 0.012	26,26 0.015	28,28 0.014	30,30 0.013				
15	3.3 0.009	7,6 0.007	6,6 0.008	7,7 0.012	8,8 0.014	10,10 0.009	11,11 0.011	12,12 0.013	13,13 0.014	15,15 0.011	16,16 0.012	17,17 0.013	18,18 0.014	19,19 0.015	21,21 0.012	22,22 0.013	23,23 0.014	24,24 0.015	26,26 0.013	27,27 0.013
20	6,4 0.005	4,4 0.008	5,5 0.009	6,6 0.010	7,7 0.011	8,8 0.011	9,9 0.011	10,10 0.011	11,11 0.011	12,12 0.011	13,13 0.011	14,14 0.012	15,15 0.012	16,16 0.012	17,17 0.012	18,18 0.012	19,19 0.012	19,19 0.015	20,20 0.015	21,21 0.015
25	4,3 0.009	7,5 0.012	4,4 0.015	5,5 0.013	6,6 0.011	7,7 0.010	8,8 0.009	9,9 0.009	9,9 0.014	10,10 0.012	11,11 0.011	12,12 0.011	12,12 0.015	13,13 0.014	14,14 0.013	15,15 0.012	16,16 0.011	16,16 0.014	17,17 0.014	18,18 0.013
30	4,3 0.006	3,3 0.012	4,4 0.009	5,5 0.007	6,6 0.006	6,6 0.012	7,7 0.010	8,8 0.008	8,8 0.013	9,9 0.011	10,10 0.009	10,10 0.013	11,11 0.011	12,12 0.010	12,12 0.013	13,13 0.012	14,14 0.011	14,14 0.014	15,15 0.012	15,15 0.015
35	2,2 0.013	3,3 0.008	4,4 0.006	4,4 0.014	5,5 0.010	6,6 0.007	6,6 0.012	7,7 0.009	7,7 0.014	8,8 0.011	9,9 0.009	9,9 0.013	10,10 0.010	10,10 0.014	11,11 0.011	11,11 0.015	12,12 0.012	13,13 0.011	13,13 0.013	14,14 0.012
40	2,2 0.008	3,3 0.008	7,5 0.013	4,4 0.007	5,5 0.006	5,5 0.012	6,6 0.008	6,6 0.013	7,7 0.009	7,7 0.013	8,8 0.010	8,8 0.014	9,9 0.011	9,9 0.014	10,10 0.011	10,10 0.014	11,11 0.012	11,11 0.014	12,12 0.012	12,12 0.014
45	2,2 0.008	6,4 0.008	3,3 0.013	4,4 0.007	4,4 0.014	5,5 0.008	5,5 0.014	6,6 0.009	6,6 0.013	7,7 0.009	7,7 0.013	8,8 0.009	8,8 0.012	9,9 0.009	9,9 0.012	10,10 0.009	10,10 0.012	10,10 0.015	11,11 0.012	11,11 0.014
50		4,3 0.013	3,3 0.010	4,4 0.005	4,4 0.010	5,5 0.006	5,5 0.010	5,5 0.015	6,6 0.009	6,6 0.013	7,7 0.009	7,7 0.012	8,8 0.009	8,8 0.011	8,8 0.014	9,9 0.011	9,9 0.013	10,10 0.010	10,10 0.012	10,10 0.015
55		4,3 0.010	3,3 0.008	7,5 0.013	4,4 0.008	4,4 0.014	5,5 0.007	5,5 0.011	6,6 0.007	6,6 0.010	6,6 0.014	7,7 0.009	7,7 0.012	8,8 0.008	8,8 0.010	8,8 0.013	9,9 0.009	9,9 0.012	9,9 0.014	10,10 0.011
60		4,3 0.008	3,3 0.007	3,3 0.014	4,4 0.006	4,4 0.011	5,5 0.006	5,5 0.009	5,5 0.013	6,6 0.007	6,6 0.010	6,6 0.014	7,7 0.009	7,7 0.011	7,7 0.014	8,8 0.010	8,8 0.012	8,8 0.015	9,9 0.010	9,9 0.013
65		4,3 0.007	3,3 0.006	3,3 0.012	6,5 0.006	4,4 0.009	4,4 0.013	5,5 0.007	5,5 0.010	5,5 0.014	6,6 0.008	6,6 0.011	6,6 0.014	7,7 0.009	7,7 0.011	7,7 0.014	8,8 0.009	8,8 0.011	8,8 0.014	9,9 0.010
70		2,2 0.014	6,4 0.008	3,3 0.010	7,5 0.013	4,4 0.007	4,4 0.011	5,5 0.005	5,5 0.008	5,5 0.011	5,5 0.015	6,6 0.008	6,6 0.011	6,6 0.014	7,7 0.009	7,7 0.011	7,7 0.013	8,8 0.009	8,8 0.011	8,8 0.013
75		2,2 0.013	4,3 0.014	3,3 0.008	3,3 0.014	4,4 0.006	4,4 0.009	4,4 0.013	5,5 0.006	5,5 0.009	5,5 0.012	6,6 0.007	6,6 0.009	6,6 0.011	6,6 0.014	7,7 0.009	7,7 0.011	7,7 0.013	8,8 0.008	8,8 0.010
80		2,2 0.011	4,3 0.012	3,3 0.007	3,3 0.012	6,5 0.006	4,4 0.008	4,4 0.011	5,5 0.005	5,5 0.007	5,5 0.010	5,5 0.013	6,6 0.007	6,6 0.009	6,6 0.012	6,6 0.014	7,7 0.008	7,7 0.010	7,7 0.013	7,7 0.015
85		2,2 0.010	4,3 0.010	3,3 0.006	3,3 0.011	7,5 0.013	4,4 0.006	4,4 0.009	4,4 0.013	5,5 0.006	5,5 0.008	5,5 0.011	5,5 0.014	6,6 0.008	6,6 0.010	6,6 0.012	6,6 0.014	7,7 0.008	7,7 0.010	7,7 0.012
90			4,3 0.009	3,3 0.005	3,3 0.009	3,3 0.014	4,4 0.005	4,4 0.008	4,4 0.011	5,5 0.005	5,5 0.007	5,5 0.009	5,5 0.012	5,5 0.015	6,6 0.008	6,6 0.010	6,6 0.012	6,6 0.014	7,7 0.008	7,7 0.010
95			4,3 0.008	6.4 0.008	3.3 0.008	3.3 0.013	6.5 0.005	4.4 0.007	4.4 0.010	4.4 0.013	5.5 0.006	5.5 0.008	5.5 0.010	5.5 0.013	6.6 0.007	6.6 0.008	6.6 0.010	6.6 0.012	6.6 0.014	7.7 0.008
100			4,3 0.007	4,3 0.014	3,3 0.007	3,3 0.011	7,5 0.013	4,4 0.006	4,4 0.008	4,4 0.011	4,4 0.015	5,5 0.007	5,5 0.009	5,5 0.011	5,5 0.013	6,6 0.007	6,6 0.009	6,6 0.010	6,6 0.012	6,6 0.014

Table A.12 Values of r, k and a for the Quantile Test when a is Approximately Equal to 0.025

Number of Site Measurements, n

Number of Reference (Background) Measurements, m

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
5	**r, k** **a**		9,9 0.030	12,12 0.024	15,15 0.021	17,17 0.026	20,20 0.024	22,22 0.028	25,25 0.025											
10		7,6 0.029	6,6 0.028	8,8 0.022	9,9 0.029	11,11 0.024	12,12 0.029	14,14 0.025	15,15 0.029	17,17 0.025	18,18 0.029	20,20 0.026	21,21 0.029	23,23 0.026	24,24 0.029	26,26 0.026	27,27 0.029			
15	11,5 0.030	6,5 0.023	5,5 0.021	6,6 0.024	7,7 0.026	8,8 0.027	9,9 0.028	10,10 0.029	11,11 0.030	13,13 0.022	14,14 0.023	15,15 0.023	16,16 0.024	17,17 0.025	18,18 0.025	19,19 0.026	21,21 0.021	21,21 0.027	22,22 0.027	23,23 0.027
20	8,4 0.023	3,3 0.030	4,4 0.026	5,5 0.024	6,6 0.022	7,7 0.020	12,11 0.021	13,12 0.024	9,9 0.028	10,10 0.026	11,11 0.024	12,12 0.023	13,13 0.022	13,13 0.029	14,14 0.027	15,15 0.026	16,16 0.025	17,17 0.024	17,17 0.029	18,18 0.028
25	2,2 0.023	8,5 0.027	6,5 0.021	7,6 0.023	5,5 0.025	6,6 0.020	10,9 0.026	7,7 0.027	8,8 0.023	13,12 0.027	9,9 0.027	10,10 0.024	11,11 0.022	11,11 0.028	12,12 0.025	13,13 0.023	13,13 0.028	14,14 0.025	15,15 0.023	15,15 0.028
30	6,3 0.026	6,4 0.026	9,6 0.026	4,4 0.021	7,6 0.029	5,5 0.026	9,8 0.024	6,6 0.029	7,7 0.023	12,11 0.021	8,8 0.025	9,9 0.021	9,9 0.027	10,10 0.023	10,10 0.029	11,11 0.025	11,11 0.030	12,12 0.026	13,13 0.023	13,13 0.027
35	7,3 0.030	4,3 0.030	3,3 0.023	6,5 0.020	4,4 0.026	10,8 0.022	5,5 0.027	9,8 0.024	6,6 0.027	7,7 0.020	7,7 0.027	8,8 0.021	8,8 0.027	9,9 0.022	9,9 0.027	10,10 0.022	10,10 0.027	11,11 0.022	11,11 0.027	12,12 0.023
40	3,2 0.029	4,3 0.022	8,5 0.028	11,7 0.025	6,5 0.028	4,4 0.030	10,8 0.026	5,5 0.027	9,8 0.023	6,6 0.026	10,9 0.028	7,7 0.024	12,11 0.020	8,8 0.023	8,8 0.029	9,9 0.022	9,9 0.027	10,10 0.021	10,10 0.026	11,11 0.021
45	3,2 0.023	8,4 0.029	6,4 0.030	3,3 0.026	8,6 0.021	4,4 0.023	7,6 0.025	5,5 0.020	5,5 0.028	9,8 0.023	6,6 0.024	10,9 0.026	7,7 0.022	7,7 0.027	8,8 0.020	8,8 0.025	8,8 0.030	9,9 0.023	9,9 0.027	10,10 0.021
50		2,2 0.025	6,4 0.022	3,3 0.021	11,7 0.027	6,5 0.026	4,4 0.026	7,6 0.028	5,5 0.021	5,5 0.028	9,8 0.022	6,6 0.023	6,6 0.029	7,7 0.020	7,7 0.025	12,11 0.020	8,8 0.022	8,8 0.026	13,12 0.027	9,9 0.023
55		2,2 0.022	4,3 0.029	8,5 0.028	3,3 0.028	8,6 0.021	4,4 0.020	4,4 0.029	10,8 0.021	5,5 0.022	5,5 0.028	9,8 0.022	6,6 0.023	6,6 0.028	10,9 0.029	7,7 0.023	7,7 0.027	12,11 0.023	8,8 0.023	8,8 0.027
60		14,5 0.022	4,3 0.024	8,5 0.021	3,3 0.023	11,7 0.029	6,5 0.024	4,4 0.023	7,6 0.023	10,8 0.024	5,5 0.023	5,5 0.029	9,8 0.022	6,6 0.022	6,6 0.027	10,9 0.027	7,7 0.021	7,7 0.025	7,7 0.030	8,8 0.021
65		6,3 0.028	7,4 0.021	6,4 0.025	10,6 0.025	3,3 0.029	8,6 0.021	6,5 0.029	4,4 0.026	7,6 0.026	10,8 0.026	5,5 0.023	5,5 0.029	9,8 0.022	6,6 0.021	6,6 0.026	10,9 0.026	7,7 0.020	7,7 0.024	7,7 0.028
70		6,3 0.024	2,2 0.029	6,4 0.021	8,5 0.028	3,3 0.025	13,8 0.026	6,5 0.023	4,4 0.022	4,4 0.028	7,6 0.028	10,8 0.027	5,5 0.024	5,5 0.029	9,8 0.022	6,6 0.021	6,6 0.025	6,6 0.029	10,9 0.030	7,7 0.022
75		11,4 0.022	2,2 0.026	4,3 0.028	8,5 0.022	3,3 0.022	9,6 0.028	8,6 0.021	6,5 0.027	4,4 0.024	7,6 0.023	7,6 0.030	10,8 0.029	5,5 0.024	5,5 0.029	9,8 0.021	6,6 0.021	6,6 0.024	6,6 0.028	10,9 0.028
80		7,3 0.028	2,2 0.024	4,3 0.024	6,4 0.028	10,6 0.024	3,3 0.027	13,8 0.027	6,5 0.023	4,4 0.020	4,4 0.026	7,6 0.024	10,8 0.023	5,5 0.020	5,5 0.025	5,5 0.029	9,8 0.021	6,6 0.020	6,6 0.024	6,6 0.027
85		3,2 0.029	2,2 0.021	4,3 0.021	6,4 0.023	8,5 0.028	3,3 0.023	9,6 0.030	8,6 0.020	6,5 0.026	4,4 0.022	4,4 0.028	7,6 0.026	10,8 0.024	5,5 0.021	5,5 0.025	5,5 0.029	9,8 0.021	6,6 0.020	6,6 0.023
90			5,3 0.020	11,5 0.027	9,5 0.023	8,5 0.023	3,3 0.021	3,3 0.028	13,8 0.028	6,5 0.022	6,5 0.029	4,4 0.024	4,4 0.029	7,6 0.028	10,8 0.026	5,5 0.022	5,5 0.025	5,5 0.030	9,8 0.021	9,8 0.025
95			10,4 0.029	2,2 0.029	4,3 0.028	6,4 0.029	10,6 0.023	3,3 0.025	11,7 0.026	8,6 0.020	6,5 0.025	4,4 0.021	4,4 0.026	7,6 0.024	7,6 0.029	10,8 0.027	5,5 0.022	5,5 0.026	5,5 0.030	9,8 0.021
100			6,3 0.029	2,2 0.027	4,3 0.025	6,4 0.025	8,5 0.028	3,3 0.022	3,3 0.029	13,8 0.028	6,5 0.022	6,5 0.028	4,4 0.023	4,4 0.027	7,6 0.025	10,8 0.022	10,8 0.028	5,5 0.022	5,5 0.026	5,5 0.030

Table A.13 Values of r, k and a for the Quantile Test when a is Approximately Equal to 0.05

Number of Site Measurements, n

Number of Reference (Background) Measurements, m

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
5	**r, k** **a**		8,8 0.051	10,10 0.057	13,13 0.043	15,15 0.048	17,17 0.051	19,19 0.054	21,21 0.056									r,k a		
10		4,4 0.043	5,5 0.057	14,12 0.045	8,8 0.046	9,9 0.052	10,10 0.058	12,12 0.046	13,13 0.050	14,14 0.054	15,15 0.057	17,17 0.049	18,18 0.052	19,19 0.055	20,20 0.057	21,21 0.059	23,23 0.053			
15	2,2 0.053	3,3 0.052	4,4 0.050	5,5 0.048	6,6 0.046	7,7 0.045	8,8 0.044	9,9 0.043	9,9 0.060	10,10 0.057	11,11 0.055	12,12 0.054	13,13 0.052	14,14 0.051	15,15 0.050	16,16 0.049	16,16 0.058	17,17 0.057	18,18 0.056	19,19 0.055
20	9,4 0.040	8,5 0.056	6,5 0.040	4,4 0.053	5,5 0.043	9,8 0.052	6,6 0.056	7,7 0.048	8,8 0.043	8,8 0.057	9,9 0.051	10,10 0.046	10,10 0.057	11,11 0.052	12,12 0.048	12,12 0.057	13,13 0.053	14,14 0.049	14,14 0.057	15,15 0.054
25	6,3 0.041	6,4 0.043	3,3 0.046	6,5 0.052	4,4 0.055	5,5 0.041	5,5 0.059	6,6 0.046	11,10 0.042	7,7 0.050	8,8 0.042	8,8 0.053	9,9 0.045	9,9 0.055	10,10 0.048	11,11 0.042	11,11 0.050	11,11 0.058	12,12 0.052	12,12 0.060
30	3,2 0.047	2,2 0.058	10,6 0.052	3,3 0.058	11,8 0.045	4,4 0.056	8,7 0.045	5,5 0.054	6,6 0.040	6,6 0.053	7,7 0.041	7,7 0.052	8,8 0.042	8,8 0.051	9,9 0.042	9,9 0.050	9,9 0.059	10,10 0.049	10,10 0.057	11,11 0.049
35	8,3 0.046	2,2 0.045	6,4 0.058	3,3 0.043	6,5 0.041	4,4 0.040	4,4 0.057	8,7 0.043	5,5 0.051	9,8 0.052	6,6 0.047	6,6 0.058	7,7 0.043	7,7 0.053	8,8 0.041	8,8 0.049	8,8 0.057	9,9 0.046	9,9 0.053	10,10 0.044
40	4,2 0.055	5,3 0.048	4,3 0.057	10,6 0.059	3,3 0.053	6,5 0.048	4,4 0.043	4,4 0.058	8,7 0.042	5,5 0.048	9,8 0.047	6,6 0.042	6,6 0.051	11,10 0.042	7,7 0.045	7,7 0.053	8,8 0.041	8,8 0.048	8,8 0.055	9,9 0.043
45	4,2 0.045	9,4 0.047	2,2 0.059	8,5 0.052	3,3 0.042	8,6 0.041	6,5 0.054	4,4 0.045	4,4 0.058	8,7 0.041	5,5 0.046	5,5 0.057	9,8 0.056	6,6 0.047	6,6 0.055	11,10 0.046	7,7 0.047	7,7 0.054	8,8 0.041	8,8 0.047
50		6,3 0.052	2,2 0.050	6,4 0.051	12,7 0.050	3,3 0.049	8,6 0.049	6,5 0.059	4,4 0.047	4,4 0.059	8,7 0.041	5,5 0.045	5,5 0.054	9,8 0.051	6,6 0.043	6,6 0.050	6,6 0.058	7,7 0.042	7,7 0.048	7,7 0.054
55		3,2 0.059	2,2 0.043	4,3 0.056	8,5 0.058	3,3 0.041	5,4 0.041	6,5 0.046	9,7 0.042	4,4 0.048	4,4 0.059	8,7 0.040	5,5 0.043	5,5 0.052	9,8 0.048	6,6 0.040	6,6 0.047	6,6 0.054	11,10 0.043	7,7 0.043
60		3,2 0.052	5,3 0.052	4,3 0.046	6,4 0.059	3,3 0.035	3,3 0.047	8,6 0.043	6,5 0.051	9,7 0.046	4,4 0.049	4,4 0.059	13,10 0.052	5,5 0.042	5,5 0.050	5,5 0.058	9,8 0.054	6,6 0.044	6,6 0.050	6,6 0.056
65		3,2 0.045	5,3 0.043	2,2 0.053	6,4 0.048	10,6 0.050	3,3 0.040	3,3 0.053	6,5 0.041	6,5 0.055	4,4 0.042	4,4 0.050	4,4 0.060	13,10 0.052	5,5 0.041	5,5 0.048	5,5 0.055	9,8 0.051	6,6 0.041	6,6 0.047
70		8,3 0.057	9,4 0.048	2,2 0.047	4,3 0.055	8,5 0.050	5,4 0.041	3,3 0.046	3,3 0.057	6,5 0.045	6,5 0.058	4,4 0.043	4,4 0.051	4,4 0.060	13,10 0.051	5,5 0.041	5,5 0.047	5,5 0.054	9,8 0.048	9,8 0.057
75		8,3 0.049	6,3 0.056	2,2 0.043	4,3 0.047	6,4 0.054	10,6 0.053	3,3 0.040	3,3 0.051	8,6 0.044	6,5 0.049	9,7 0.041	4,4 0.044	4,4 0.052	5,5 0.060	13,10 0.051	8,7 0.047	5,5 0.046	5,5 0.052	5,5 0.058
80		4,2 0.059	6,3 0.048	5,3 0.053	2,2 0.055	6,4 0.046	8,5 0.055	5,4 0.042	3,3 0.045	3,3 0.055	6,5 0.041	6,5 0.052	9,7 0.043	4,4 0.045	4,4 0.053	7,6 0.058	13,10 0.051	8,7 0.046	5,5 0.045	5,5 0.051
85		4,2 0.054	3,2 0.058	5,3 0.047	2,2 0.050	4,3 0.054	4,3 0.048	10,6 0.056	5,4 0.049	3,3 0.049	3,3 0.059	6,5 0.044	6,5 0.055	9,7 0.046	4,4 0.046	4,4 0.053	7,6 0.059	10,8 0.060	8,7 0.045	5,5 0.044
90			3,2 0.053	5,3 0.041	2,2 0.046	6,4 0.059	6,4 0.051	8,5 0.058	5,4 0.042	3,3 0.044	3,3 0.053	8,6 0.045	6,5 0.047	6,5 0.058	4,4 0.041	4,4 0.047	4,4 0.054	7,6 0.059	10,8 0.060	8,7 0.045
95			3,2 0.048	9,4 0.048	2,2 0.042	2,2 0.056	4,3 0.059	8,5 0.050	10,6 0.058	5,4 0.048	3,3 0.048	3,3 0.056	6,5 0.041	6,5 0.050	9,7 0.040	4,4 0.042	4,4 0.048	4,4 0.054	7,6 0.059	10,8 0.059
100			3,2 0.044	6,3 0.057	5,3 0.054	2,2 0.052	4,3 0.053	6,4 0.056	10,6 0.049	5,4 0.043	3,3 0.043	3,3 0.051	3,3 0.059	6,5 0.044	6,5 0.053	9,7 0.042	4,4 0.043	4,4 0.049	4,4 0.055	7,6 0.059

Table A.14 Values of r, k and a for the Quantile Test when a is Approximately Equal to 0.10

Number of Site Measurements, n

Number of Reference (Background) Measurements, m

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
5	**r, k** **a**		7,7 0.083	8,8 0.116	10,10 0.109	12,12 0.104	14,14 0.100	15,15 0.117	17,17 0.112									r,k a		
10		3,3 0.105	4,4 0.108	5,5 0.109	6,6 0.109	7,7 0.109	8,8 0.109	9,9 0.109	10,10 0.109	11,11 0.109	12,12 0.109	13,13 0.109	14,14 0.109	15,15 0.109	16,16 0.109	17,17 0.109	18,18 0.109			
15	9,4 0.098	10,6 0.106	3,3 0.112	4,4 0.093	5,5 0.081	5,5 0.117	6,6 0.102	7,7 0.092	7,7 0.118	8,8 0.106	9,9 0.098	9,9 0.118	10,10 0.109	11,11 0.101	11,11 0.118	12,12 0.110	13,13 0.104	13,13 0.118	14,14 0.111	15,15 0.106
20	3,2 0.091	2,2 0.103	5,4 0.093	3,3 0.115	4,4 0.085	4,4 0.119	5,5 0.093	10,9 0.084	6,6 0.099	7,7 0.083	7,7 0.102	8,8 0.088	8,8 0.105	9,9 0.092	9,9 0.107	10,10 0.095	10,10 0.108	11,11 0.098	11,11 0.110	12,12 0.100
25	4,2 0.119	7,4 0.084	8,5 0.112	3,3 0.080	3,3 0.117	4,4 0.080	4,4 0.107	8,7 0.108	5,5 0.101	10,9 0.088	6,6 0.096	6,6 0.114	7,7 0.093	7,7 0.108	8,8 0.091	8,8 0.104	8,8 0.117	9,9 0.100	9,9 0.112	10,10 0.098
30	4,2 0.089	5,3 0.089	2,2 0.106	14,8 0.111	3,3 0.088	3,3 0.119	9,7 0.116	4,4 0.100	8,7 0.093	5,5 0.088	5,5 0.106	6,6 0.080	6,6 0.095	6,6 0.110	7,7 0.087	7,7 0.100	7,7 0.113	8,8 0.092	8,8 0.103	8,8 0.115
35	5,2 0.109	3,2 0.119	2,2 0.086	6,4 0.119	5,4 0.091	3,3 0.093	3,3 0.120	9,7 0.112	4,4 0.094	4,4 0.114	8,7 0.107	5,5 0.094	5,5 0.110	6,6 0.081	6,6 0.094	6,6 0.107	6,6 0.120	7,7 0.094	7,7 0.105	7,7 0.116
40	5,2 0.087	3,2 0.098	5,3 0.119	2,2 0.107	12,7 0.109	5,4 0.102	3,3 0.097	6,5 0.100	9,7 0.109	4,4 0.090	4,4 0.107	8,7 0.097	5,5 0.086	5,5 0.099	5,5 0.112	6,6 0.082	6,6 0.093	6,6 0.104	6,6 0.116	7,7 0.089
45	6,2 0.103	3,2 0.082	5,3 0.094	2,2 0.091	6,4 0.115	7,5 0.086	5,4 0.112	3,3 0.100	6,5 0.101	9,7 0.107	4,4 0.087	4,4 0.102	4,4 0.117	8,7 0.107	5,5 0.091	5,5 0.103	5,5 0.115	6,6 0.083	6,6 0.093	6,6 0.103
50		7,3 0.083	9,4 0.115	7,4 0.097	2,2 0.108	10,6 0.112	5,4 0.090	3,3 0.084	3,3 0.103	6,5 0.102	9,7 0.105	4,4 0.084	4,4 0.098	4,4 0.112	8,7 0.099	5,5 0.084	5,5 0.095	5,5 0.105	5,5 0.116	6,6 0.083
55		4,2 0.109	3,2 0.114	5,3 0.114	2,2 0.095	6,4 0.112	14,8 0.111	5,4 0.098	3,3 0.088	3,3 0.105	6,5 0.103	9,7 0.104	4,4 0.082	4,4 0.095	4,4 0.107	4,4 0.120	8,7 0.107	5,5 0.088	5,5 0.098	5,5 0.108
60		4,2 0.095	3,2 0.100	5,3 0.097	2,2 0.084	2,2 0.109	8,5 0.119	5,4 0.082	5,4 0.105	3,3 0.091	3,3 0.106	6,5 0.103	9,7 0.102	4,4 0.081	4,4 0.092	4,4 0.103	4,4 0.115	8,7 0.100	5,5 0.083	5,5 0.092
65		4,2 0.084	3,2 0.089	5,3 0.082	7,4 0.090	2,2 0.097	6,4 0.110	12,7 0.113	5,4 0.089	5,4 0.111	3,3 0.093	3,3 0.108	6,5 0.104	9,7 0.101	7,6 0.084	4,4 0.090	4,4 0.100	4,4 0.110	8,7 0.094	8,7 0.107
70		5,2 0.115	7,3 0.101	9,4 0.106	5,3 0.112	2,2 0.088	2,2 0.109	8,5 0.114	7,5 0.081	5,4 0.096	3,3 0.083	3,3 0.096	3,3 0.109	6,5 0.104	9,7 0.101	7,6 0.082	4,4 0.088	4,4 0.097	4,4 0.107	4,4 0.117
75		5,2 0.103	7,3 0.088	3,2 0.111	5,3 0.098	7,4 0.101	2,2 0.099	2,2 0.119	10,6 0.117	5,4 0.083	5,4 0.102	3,3 0.085	3,3 0.098	3,3 0.110	6,5 0.105	9,7 0.100	7,6 0.081	4,4 0.086	4,4 0.095	4,4 0.104
80		5,2 0.093	4,2 0.116	3,2 0.101	5,3 0.086	7,4 0.086	2,2 0.091	2,2 0.109	8,5 0.110	14,8 0.110	5,4 0.089	5,4 0.107	3,3 0.088	3,3 0.099	3,3 0.111	6,5 0.105	6,5 0.120	9,7 0.116	4,4 0.084	4,4 0.093
85		4,2 0.106	4,2 0.106	3,2 0.092	9,4 0.117	5,3 0.111	2,2 0.083	2,2 0.101	2,2 0.118	10,6 0.112	7,5 0.084	5,4 0.094	5,4 0.111	3,3 0.090	3,3 0.101	3,3 0.112	6,5 0.105	6,5 0.119	9,7 0.114	4,4 0.083
90			4,2 0.097	3,2 0.085	3,2 0.119	5,3 0.099	7,4 0.095	2,2 0.093	2,2 0.109	8,5 0.108	12,7 0.114	5,4 0.083	5,4 0.099	3,3 0.082	3,3 0.092	3,3 0.102	3,3 0.112	6,5 0.105	6,5 0.119	9,7 0.113
95			4,2 0.089	7,3 0.100	3,2 0.110	5,3 0.089	7,4 0.084	2,2 0.086	2,2 0.102	2,2 0.117	10,6 0.108	14,8 0.117	5,4 0.088	5,4 0.103	3,3 0.084	3,3 0.094	3,3 0.103	3,3 0.113	6,5 0.106	6,5 0.118
100			4,2 0.082	7,3 0.090	3,2 0.102	5,3 0.080	5,3 0.109	2,2 0.080	2,2 0.095	2,2 0.110	6,4 0.118	12,7 0.109	7,5 0.086	5,4 0.093	5,4 0.108	3,3 0.086	3,3 0.095	3,3 0.104	3,3 0.114	6,5 0.106

Table A.15 Critical Values, w_a, for the Wilcoxon Rank Sum (WRS) Test. (n = the Number of Site Measurements; m = the Number of Background Measurements)

n	a	m																		
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**191**	**20**
2	0.05	0	0	0	1	1	1	2	2	2	2	3	3	4	4	4	4	5	5	5
	0.10	0	1	1	2	2	2	3	3	4	4	5	5	5	6	6	7	7	8	8
3	0.05	0	1	1	2	3	3	4	5	5	6	6	7	8	8	9	10	10	11	12
	0.10	1	2	2	3	4	5	6	6	7	8 1	9	10	11	11	12	13	14	15	16
4	0.05	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19
	0.10	1	2	4	5	6	7	8	10	11	12	13	14	16	17	18	19	21	22	23
5	0.05	1	2	3	5	6	7	9	10	12	13	14	16	17	19	20	21	23	24	26
	0.10	2	3	5	6	8	9	11	13	14	16	18	19	21	23	24	26	28	29	31
6	0.05	1	3	4	6	8	9	11	13	15	17	18	20	22	24	26	27	29	31	33
	0.10	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	35	37	39
7	0.05	1	3	5	7	9	12	14	16	18	20	22	25	27	29	31	34	36	38	40
	0.10	2	5	7	9	12	14	17	19	22	24	27	29	32	34	37	39	42	44	47
8	0.05	2	4	6	9	11	14	16	19	21	24	27	29	32	34	37	40	42	45	48
	0.10	3	6	8	11	14	17	20	23	25	28	31	34	37	40	43	46	49	52	55
9	0.05	2	5	7	10	13	16	19	22	25	28	31	34	37	40	43	46	49	52	55
	0.10	3	6	10	13	16	19	23	26	29	32	36	39	42	46	49	53	56	59	63
10	0.05	2	5	8	12	15	18	21	25	28	32	35	38	42	45	49	52	56	59	63
	0.10	4	7	11	14	18	22	25	29	33	37	40	44	48	52	55	59	63	67	71
11	0.05	2	6	9	13	17	20	24	28	32	35	39	43	47	51	55	58	62	66	70
	0.10	4	8	12	16	20	24	28	32	37	41	45	49	53	58	62	66	70	74	79
12	0.05	3	6	10	14	18	22	27	31	35	39	43	48	52	56	61	65	69	73	78
	0.10	5	9	13	18	22	27	31	36	40	45	50	54	59	64	68	73	78	82	87
13	0.05	3	7	11	16	20	25	29	34	38	43	48	52	57	62	66	71	76	81	8
	0.10	5	10	14	19	24	29	34	39	44	49	54	59	64	69	75	80	85	90	95
14	0.05	4	8	12	17	22	27	32	37	42	47	52	57	62	67	72	78	83	88	93
	0.10	5	11	16	21	26	32	37	42	48	53	59	64	70	75	81	86	92	98	103
15	0.05	4	8	13	19	24	29	34	40	45	51	56	62	67	73	78	84	89	95	101
	0.10	6	11	17	23	28	34	40	46	52	58	64	69	75	81	87	93	99	105	111
16	0.05	4	9	15	20	26	31	37	43	49	55	61	66	72	78	84	90	96	102	108
	0.10	6	12	18	24	30	37	43	49	55	62	68	75	81	87	94	100	107	113	120
17	0.05	4	10	16	21	27	34	40	46	52	58	65	71	78	84	90	97	103	110	116
	0.10	7	13	19	26	32	39	46	53	59	66	73	80	86	93	100	107	114	121	128
18	0.05	5	10	17	23	29	36	42	49	56	62	69	76	83	89	96	103	110	117	124
	0.10	7	14	21	28	35	42	49	56	63	70	78	85	92	99	107	114	121	129	136
19	0.05	5	11	18	24	31	38	45	52	59	66	73	81	88	95	102	110	117	124	131
	0.10	8	15	22	29	37	44	52	59	67	74	82	90	98	105	113	121	129	136	144

n	a	m																	
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**191**	**20**
20	0.05 0.10	5 8	12 16	19 23	26 31	33 39	40 47	48 55	55 63	63 71	70 79	78 87	85 95	93 103	101 111	108 120	116 128	124 136	131 144	139 152

Table A.16. Critical Values for the Two-Sample t Test

Degrees of Freedom	1 - a								
	.70	.75	.80	.85	.90	.95	.975	.99	.995
1	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.536	0.691	0.866	1.074	1.34	1.753	2.131	2.602	2.947
16	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
40	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704
60	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660
120	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617
∞	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576