United States Environmental Protection Agency

Office of Research and Development Washington, D.C. 20460

EPA/600/R-96/084 July 1996

GUIDANCE FOR DATA QUALITY ASSESSMENT

Practical Methods for Data Analysis

EPA QA/G-9

QA96 Version

FOREWORD

This is the 1996 (QA96) version of *Guidance for Data Quality Assessment*, EPA QA/G-9. The Environmental Protection Agency (EPA) has developed the Data Quality Assessment (DQA) Process as an important tool for project managers and planners to determine whether the type, quantity, and quality of data needed to support Agency decisions has been achieved. This guidance is the culmination of experiences in the design and statistical analyses of environmental data in different Program Offices at the EPA. Many elements of prior guidance, statistics, and scientific planning have been incorporated into this document.

This document provides general guidance to organizations on assessing data quality criteria and performance specifications for decision making. This guidance assumes that an appropriate Quality System has been established and that planning for data collection has been achieved using a scientifically-based information collection strategy. An overview of the Agency's recommended data collection procedure, the DQO Process, is included in this guidance in Chapter 1 and EPA QA/G-4.

Guidance for Data Quality Assessment is distinctly different from other guidance documents; it is not intended to be read in a linear or continuous fashion. The intent of the document is for it to be used as a "tool-box" of useful techniques in assessing the quality of data. The overall structure of the document will enable the analyst to investigate many different problems using a systematic methodology. The methodology consists of five steps that should be iterated between them as necessary:

(i) Review the Data Quality Objectives
(ii) Conduct a Preliminary Data Review
(iii) Select the Statistical Test
(iv) Verify the Assumptions of the Test
(v) Draw Conclusions From the Data

This approach closely parallels the activities of a statistician analyzing a data set for the first time. The five step procedure is not intended to be a definitive analysis of a project or problem, but provide an initial assessment on the "reasonableness" of the data that have been generated. Sophisticated statistical analysis is often not necessary unless special or unusual circumstances have been encountered in the generation or collection of the data or the analysis is planned in detail before the data are collected. This guidance is directed towards the analysis of relatively small data sets containing data that have been collected in a relatively simple fashion. The analysis of survey data containing large data sets or a complex sampling scheme is best left for statistical experts.

This document is a product of the collaborative effort of many quality management professionals throughout the EPA and the environmental community. It has been peer reviewed by the EPA Program Offices, Regional Offices, and Laboratories. Many valuable comments and suggestions have been incorporated to make it more useful, and additional suggestions to improve its effectiveness are sought. The Quality Assurance Division has the Agency lead for the development of statistical quality assurance techniques and future editions of this guidance will contain some of these recent developments.

This document is one of a series of quality management guidance documents that the EPA Quality Assurance Division (QAD) has prepared to assist users in implementing the Agency-wide Quality System. Other related documents currently available or planned include:

EPA QA/G-4 *Guidance for The Data Quality Objectives Process*

EPA QA/G-4D *DEFT Software for the Data Quality Objectives Process*

EPA QA-G-4R *Guidance for the Data Quality Objectives Process for Researchers (planned)*

EPA QA/G-4S *Guidance for the Data Quality Objectives Process (Superfund)*

EPA QA/G-5 *Guidance for Quality Assurance Project Plans (draft)*

EPA QA/G-5S *Guidance on Sampling Plans (planned)*

EPA QA/G-6 *Guidance for the Preparation of Standard Operating Procedures (SOPs) for Quality-Related Documents*

EPA QA/G-9D *Data Quality Evaluation Statistical Tools (DataQUEST)*

The External Comment Draft EPA QA/G-5, the Final Version of EPA QA/G-4S, and the External Comment Draft EPA of QA/G-4R and QA/G-5S should be available before December 1996.

This document is intended to be a "living document" that will be updated annually to incorporate new topics and revisions or refinements to existing procedures. Comments received on this 1996 version will be considered for inclusion in subsequent versions. In addition, user-friendly PC-based software (EPA QA/G-9D) to supplement this guidance is being developed and should be available from QAD in September 1996.

Please send your written comments on *Guidance for Data Quality Assessment* to:

Quality Assurance Division (8724)
Office of Research and Development
U.S. Environmental Protection Agency
401 M Street, SW
Washington, DC 20460
(202) 260-5763
FAX (202) 401-7002
E-mail: ord-qad@epamail.epa.gov

TABLE OF CONTENTS

	Page
INTRODUCTION	0 - 1
0.1 PURPOSE AND OVERVIEW	0 - 1
0.2 DQA AND THE DATA LIFE CYCLE	0 - 2
0.3 THE 5 STEPS OF THE DQA PROCESS	0 - 2
0.4 INTENDED AUDIENCE	0 - 3
0.5 ORGANIZATION	0 - 4
0.6 SUPPLEMENTAL SOURCES	0 - 4
0.7 SCOPE AND LIMITATIONS	0 - 4
STEP 1: REVIEW DQOs AND THE SAMPLING DESIGN	1.1 - 1
1.1 OVERVIEW AND ACTIVITIES	1.1 - 1
1.1.1 Review Study Objectives	1.1 - 1
1.1.2 Translate Objectives into Statistical Hypotheses	1.1 - 2
1.1.3 Develop Limits on Decision Errors	1.1 - 2
1.1.4 Review Sampling Design	1.1 - 3
1.2 DEVELOPING THE STATEMENT OF HYPOTHESES	1.2 - 1
1.3 DESIGNS FOR SAMPLING ENVIRONMENTAL MEDIA	1.3 - 1
1.3.1 Authoritative Sampling	1.3 - 1
1.3.2 Probability Sampling	1.3 - 1
1.3.2.1 Simple Random Sampling	1.3 - 1
1.3.2.2 Sequential Random Sampling	1.3 - 2
1.3.2.3 Systematic Samples	1.3 - 2
1.3.2.4 Stratified Samples	1.3 - 2
1.3.2.5 Compositing Physical Samples	1.3 - 3
1.3.2.6 Other Sampling Designs	1.3 - 3
STEP 2: CONDUCT A PRELIMINARY DATA REVIEW	2.1 - 1
2.1 OVERVIEW AND ACTIVITIES	2.1 - 1
2.1.1 Review Quality Assurance Reports	2.1 - 1
2.1.2 Calculate Basic Statistical Quantities	2.1 - 2
2.1.3 Graph the Data	2.1 - 2
2.2 STATISTICAL QUANTITIES	2.2 - 1
2.2.1 Measures of Relative Standing	2.2 - 1
2.2.2 Measures of Central Tendency	2.2 - 2
2.2.3 Measures of Dispersion	2.2 - 2
2.2.4 Measures of Association	2.2 - 5
2.3 GRAPHICAL REPRESENTATIONS	2.3 - 1
2.3.1 Histogram/Frequency Plots	2.3 - 1
2.3.2 Stem-and-Leaf Plot	2.3 - 3
2.3.3 Box and Whisker Plot	2.3 - 3
2.3.4 Ranked Data Plot	2.3 - 6
2.3.5 Quantile Plot	2.3 - 8
2.3.6 Normal Probability Plot (Quantile-Quantile Plot)	2.3 - 10

Page

2.3.7 Plots for Two or More Variables . . . . . 2.3 - 13
2.3.7.1 Plots for Individual Data Points . . . . . 2.3 - 13
2.3.7.2 Scatter Plot . . . . . 2.3 - 14
2.3.7.3 Extensions of the Scatter Plot . . . . . 2.3 - 15
2.3.7.4 Empirical Quantile-Quantile Plot . . . . . 2.3 - 16
2.3.8 Plots for Temporal Data . . . . . 2.3 - 18
2.3.8.1 Time Plot . . . . . 2.3 - 19
2.3.8.2 Plot of the Autocorrelation Function (Correlogram) . . . . . 2.3 - 20
2.3.8.3 Multiple Observations Per Time Period . . . . . 2.3 - 22
2.3.9 Plots for Spatial Data . . . . . 2.3 - 23
2.3.9.1 Posting Plots . . . . . 2.3 - 23
2.3.9.2 Symbol Plots . . . . . 2.3 - 23
2.3.9.3 Other Spatial Graphical Representations . . . . . 2.3 - 25

STEP 3: SELECT THE STATISTICAL TEST . . . . . 3.1 - 1
3.1 OVERVIEW AND ACTIVITIES . . . . . 3.1 - 1
3.1.1 Select Statistical Hypothesis Test . . . . . 3.1 - 1
3.1.2 Identify Assumptions Underlying the Statistical Test . . . . . 3.1 - 1
3.2 TESTS OF HYPOTHESES ABOUT A SINGLE POPULATION . . . . . 3.2 - 1
3.2.1 Tests for a Mean . . . . . 3.2 - 1
3.2.1.1 The One-Sample t-Test . . . . . 3.2 - 2
3.2.1.2 The Wilcoxon Signed Rank (One-Sample) Test for the Mean . . . . . 3.2 - 7
3.2.2 Tests for a Proportion or Percentile . . . . . 3.2 - 11
3.2.2.1 The One-Sample Proportion Test . . . . . 3.2 - 11
3.2.3 Tests for a Median . . . . . 3.2 - 13
3.3 TESTS FOR COMPARING TWO POPULATIONS . . . . . 3.3 - 1
3.3.1 Comparing Two Means . . . . . 3.3 - 1
3.3.1.1 Student's Two-Sample t-Test (Equal Variances) . . . . . 3.3 - 2
3.3.1.2 Satterthwaite's Two-Sample t-Test (Unequal Variances) . . . . . 3.3 - 2
3.3.2 Comparing Two Proportions or Percentiles . . . . . 3.3 - 7
3.3.2.1 Two-Sample Test for Proportions . . . . . 3.3 - 7
3.3.3 Nonparametric Comparisons of Two Population . . . . . 3.3 - 10
3.3.3.1 The Wilcoxon Rank Sum Test . . . . . 3.3 - 10
3.3.3.2 The Quantile Test . . . . . 3.3 - 14
3.3.4 Comparing Two Medians . . . . . 3.3 - 14

STEP 4: VERIFY THE ASSUMPTIONS OF THE STATISTICAL TEST . . . . . 4.1 - 1
4.1 OVERVIEW AND ACTIVITIES . . . . . 4.1 - 1
4.1.1 Determine Approach for Verifying Assumptions . . . . . 4.1 - 1
4.1.2 Perform Tests of Assumptions . . . . . 4.1 - 2
4.1.3 Determine Corrective Actions . . . . . 4.1 - 2
4.2 TESTS FOR DISTRIBUTIONAL ASSUMPTIONS . . . . . 4.2 - 1
4.2.1 Graphical Methods . . . . . 4.2 - 3
4.2.2 Shapiro-Wilk Test for Normality (the W test) . . . . . 4.2 - 3
4.2.3 Extensions of the Shapiro-Wilk Test (Filliben's Statistic) . . . . . 4.2 - 3
4.2.4 Coefficient of Variation . . . . . 4.2 - 4
4.2.5 Coefficient of Skewness/Coefficient of Kurtosis Tests . . . . . 4.2 - 4

Page

4.2.6 Range Tests . . . . . 4.2 - 5
4.2.7 Goodness-of-Fit Tests . . . . . 4.2 - 7
4.2.8 Recommendations . . . . . 4.2 - 7
4.3 TESTS FOR TRENDS . . . . . 4.3 - 1
4.3.1 Introduction . . . . . 4.3 - 1
4.3.2 Regression-Based Methods for Estimating and Testing for Trends . . . . . 4.3 - 1
4.3.2.1 Estimating a Trend Using the Slope of the Regression Line . . . . . 4.3 - 1
4.3.2.2 Testing for Trends Using Regression Methods . . . . . 4.3 - 2
4.3.3 General Trend Estimation Methods . . . . . 4.3 - 3
4.3.3.1 Sen's Slope Estimator . . . . . 4.3 - 3
4.3.3.2 Seasonal Kendall Slope Estimator . . . . . 4.3 - 3
4.3.4 Hypothesis Tests for Detecting Trends . . . . . 4.3 - 3
4.3.4.1 One Observation per Time Period for One Sampling Location . . . . . 4.3 - 3
4.3.4.2 Multiple Observations per Time Period for One Sampling Location . . . . . 4.3 - 7
4.3.4.3 Multiple Sampling Locations with Multiple Observations . . . . . 4.3 - 7
4.3.4.4 One Observation for One Station with Multiple Seasons . . . . . 4.3 - 9
4.3.5 A Discussion on Tests for Trends . . . . . 4.3 - 10
4.4 OUTLIERS . . . . . 4.4 - 1
4.4.1 Background . . . . . 4.4 - 1
4.4.2 Selection of a Statistical Test . . . . . 4.4 - 2
4.4.3 Extreme Value Test (Dixon's Test) . . . . . 4.4 - 2
4.4.4 Discordance Test . . . . . 4.4 - 4
4.4.5 Rosner's Test . . . . . 4.4 - 5
4.4.6 Walsh's Test . . . . . 4.4 - 7
4.4.7 Multivariate Outliers . . . . . 4.4 - 7
4.5 TESTS FOR DISPERSIONS . . . . . 4.5 - 1
4.5.1 Confidence Intervals for a Single Variance . . . . . 4.5 - 1
4.5.2 The F-Test for the Equality of Two Variances . . . . . 4.5 - 1
4.5.3 Bartlett's Test for the Equality of Two or More Variances . . . . . 4.5 - 1
4.5.4 Levene's Test for the Equality of Two or More Variances . . . . . 4.5 - 4
4.6 TRANSFORMATIONS . . . . . 4.6 - 1
4.6.1 Types of Data Transformations . . . . . 4.6 - 1
4.6.2 Reasons for Data Transformations . . . . . 4.6 - 2
4.7 VALUES BELOW DETECTION LIMITS . . . . . 4.7 - 1
4.7.1 Less than 15% Nondetects - Substitution Methods . . . . . 4.7 - 2
4.7.2 Between 15-50% Nondetects . . . . . 4.7 - 2
4.7.2.1 Cohen's Method . . . . . 4.7 - 2
4.7.2.2 Trimmed Mean . . . . . 4.7 - 4
4.7.2.3 Winsorized Mean and Standard Deviation . . . . . 4.7 - 5
4.7.3 Greater than 50% Nondetects - Test of Proportions . . . . . 4.7 - 6

STEP 5: DRAW CONCLUSIONS FROM THE DATA . . . . . 5.1 - 1
5.1 OVERVIEW AND ACTIVITIES . . . . . 5.1 - 1
5.1.1 Perform the Statistical Hypothesis Test . . . . . 5.1 - 1
5.1.2 Draw Study Conclusions . . . . . 5.1 - 1
5.1.3 Evaluate Performance of the Sampling Design . . . . . 5.1 - 2

Page

5.2 INTERPRETING AND COMMUNICATING THE TEST RESULTS . . . . . 5.2 - 1
5.2.1 Interpretation of p-Values . . . . . 5.2 - 1
5.2.2 "Accepting" vs. "Failing to Reject" the Null Hypothesis . . . . . 5.2 - 1
5.2.3 Statistical Significance vs. Practical Significance . . . . . 5.2 - 2
5.2.4 Impact of Bias on Test Results . . . . . 5.2 - 2
5.2.5 Quantity vs. Quality of Data . . . . . 5.2 - 5
5.2.6 "Proof of Safety" vs. "Proof of Hazard" . . . . . 5.2 - 6

LIST OF APPENDICES

Page

A. STATISTICAL TABLES . . . . . A - 1
B. REFERENCES . . . . . B - 1

LIST OF FIGURES

Figure No. Page

0.2-1. DQA in the Context of the Data Life Cycle . . . . . 0 - 2
2.3-1. Example of a Frequency Plot . . . . . 2.3 - 1
2.3-2. Example of a Histogram . . . . . 2.3 - 1
2.3-3. Example of a Box and Whisker Plot . . . . . 2.3 - 3
2.3-4. Example of a Ranked Data Plot . . . . . 2.3 - 6
2.3-5. Example of a Quantile Plot of Skewed Data . . . . . 2.3 - 8
2.3-6. Normal Probability Paper . . . . . 2.3 - 12
2.3-7. Example of Graphical Representations of Multiple Variables . . . . . 2.3 - 13
2.3-8. Example of a Scatter Plot . . . . . 2.3 - 14
2.3-9. Example of a Coded Scatter Plot . . . . . 2.3 - 15
2.3-10. Example of a Parallel Coordinates Plot . . . . . 2.3 - 15
2.3-11. Example of a Matrix Scatter Plot . . . . . 2.3 - 16
2.3.12. Example of a Time Plot Showing a Slight Downward Trend . . . . . 2.3 - 19
2.3-13. Example of a Correlogram . . . . . 2.3 - 20
2.3-14. Example of a Posting Plot . . . . . 2.3 - 23
2.3-15. Example of a Symbol Plot . . . . . 2.3 - 24
4.2-1. Graph of a Normal and Lognormal Distribution . . . . . 4.2 - 1
5.2-1. Illustration of Unbiased versus Biased Power Curves . . . . . 5.2 - 5

LIST OF TABLES

Table No. Page

1.2-1. Commonly Used Statements of Statistical Hypotheses . . . . . 1.2 - 3
4.2-1. Data for Examples in Section 4.2 . . . . . 4.2 - 1
4.2-2. Tests for Normality . . . . . 4.2 - 2
4.4-1. Recommendations for Selecting a Statistical Test for Outliers . . . . . 4.4 - 2
4.7-1. Guidelines for Analyzing Data with Nondetects . . . . . 4.7 - 1

INTRODUCTION

0.1 PURPOSE AND OVERVIEW

Data Quality Assessment (DQA) 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. This guidance demonstrates how to use DQA in evaluating environmental data sets and illustrates how to apply some graphical and statistical tools for performing DQA. The guidance focuses primarily on using DQA in environmental decision making; however, the tools presented for preliminary data review and verifying statistical assumptions are useful whenever environmental data are used, regardless of whether the data are used for decision making.

DQA is built on a fundamental premise: data *quality*, as a concept, is meaningful only when it relates to the *intended use* of the data. Data quality does not exist in a vacuum; one must know in what context a data set is to be used in order to establish a relevant yardstick for judging whether or not the data set is adequate. By using the DQA Process, one can answer two fundamental questions:

1. Can the decision (or estimate) be made with the desired confidence, given the quality of the data set?

2. How well can the sampling design be expected to perform over a wide range of possible outcomes? If the same sampling design strategy is used again for a similar study, would the data be expected to support the same intended use with the desired level of confidence, particularly if the measurement results turned out to be higher or lower than those observed in the current study?

The first question addresses the data user's immediate needs. For example, if the data provide evidence strongly in favor of one course of action over another, then the decision maker can proceed knowing that the decision will be supported by unambiguous data. If, however, the data do not show sufficiently strong evidence to favor one alternative, then the data analysis alerts the decision maker to this uncertainty. The decision maker now is in a position to make an informed choice about how to proceed (such as collect more or different data before making the decision, or proceed with the decision despite the relatively high, but acceptable, probability of drawing an erroneous conclusion).

The second question addresses the data user's potential future needs. For example, if investigators decide to use a certain sampling design at a different location from where the design was first used, they should determine how well the design can be expected to perform given that the outcomes and environmental conditions of this sampling event will be different from those of the original event. Because environmental conditions will vary from one location or time to another, the adequacy of the sampling design approach should be evaluated over a broad range of possible outcomes and conditions.

0.2 DQA AND THE DATA LIFE CYCLE

The data life cycle (depicted in Figure 0.2-1) comprises three steps: planning, implementation, and assessment. During the planning phase, the Data Quality Objectives (DQO) Process (or some other systematic planning procedure) is used to define quantitative and qualitative criteria for determining when, where, and how many samples (measurements) to collect and a desired level of confidence. This information, along with the sampling methods, analytical procedures, and appropriate quality assurance (QA) and quality

control (QC) procedures, are documented in the Quality Assurance Project Plan (QAPP). Data are then collected following the QAPP specifications. DQA completes the data life cycle by providing the assessment needed to determine if the planning objectives were achieved. During the assessment phase, the data are validated and verified to ensure that the sampling and analysis protocols specified in the QAPP were followed, and that the measurement systems performed in accordance with the criteria specified in the QAPP. DQA then proceeds using the validated data set to determine if the quality of the data is satisfactory.

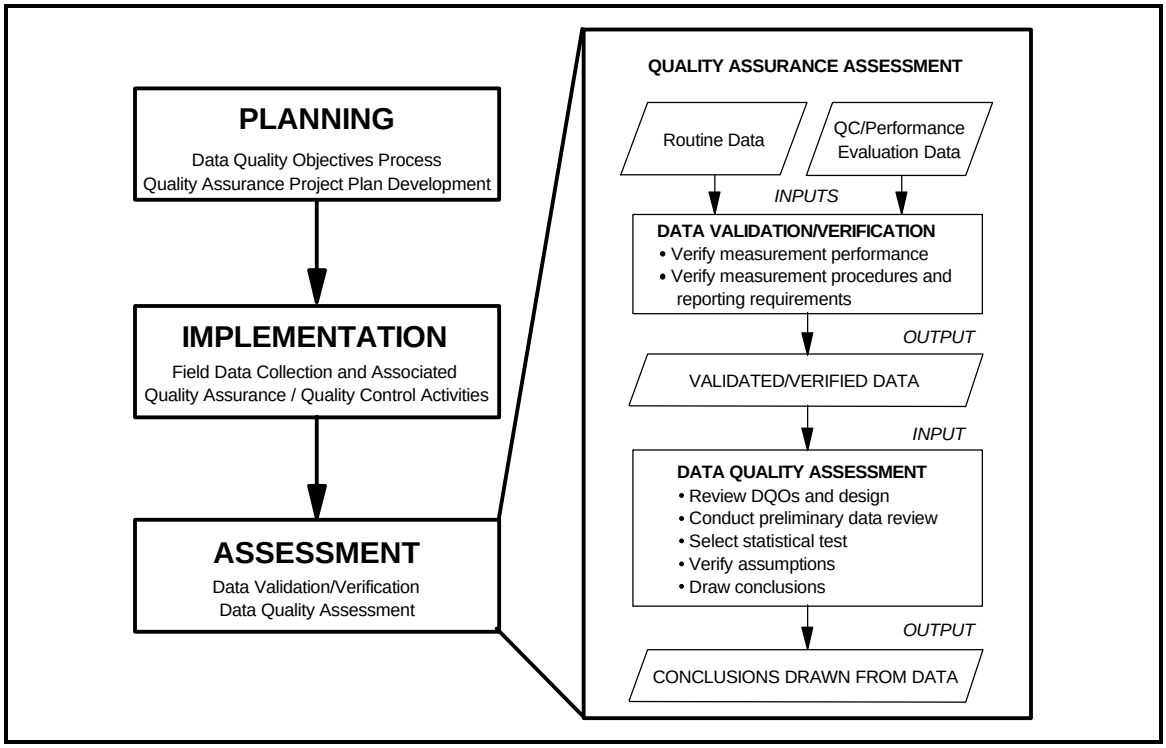


Figure 0.2-1. DQA in the Context of the Data Life Cycle

0.3 THE 5 STEPS OF THE DQA PROCESS

The DQA Process involves five steps that begin with a review of the planning documentation and end with an answer to the question posed during the planning phase of the study. These steps roughly parallel the actions of an environmental statistician when analyzing a set of data. The five steps, which are described in detail in the remaining chapters of this guidance, are briefly summarized as follows:

1. ***Review the Data Quality Objectives (DQOs) and Sampling Design:*** Review the DQO outputs to assure that they are still applicable. If DQOs have not been developed, specify DQOs before evaluating the data (e.g., for environmental decisions, define the statistical hypothesis and specify tolerable limits on decision errors; for estimation problems, define an acceptable confidence or probability interval width). Review the sampling design and data collection documentation for consistency with the DQOs.

2. ***Conduct a Preliminary Data Review:*** Review QA reports, calculate basic statistics, and generate graphs of the data. Use this information to learn about the structure of the data and identify patterns, relationships, or potential anomalies.

3. ***Select the Statistical Test:*** Select the most appropriate procedure for summarizing and analyzing the data, based on the review of the DQOs, the sampling design, and the preliminary data review. Identify the key underlying assumptions that must hold for the statistical procedures to be valid.

4. ***Verify the Assumptions of the Statistical Test:*** Evaluate whether the underlying assumptions hold, or whether departures are acceptable, given the actual data and other information about the study.

5. ***Draw Conclusions from the Data:*** Perform the calculations required for the statistical test and document the inferences drawn as a result of these calculations. If the design is to be used again, evaluate the performance of the sampling design.

These five steps are presented in a linear sequence, but the DQA process is by its very nature iterative. For example, if the preliminary data review reveals patterns or anomalies in the data set that are inconsistent with the DQOs, then some aspects of the study planning may have to be reconsidered in Step 1. Likewise, if the underlying assumptions of the statistical test are not supported by the data, then previous steps of the DQA process may have to be revisited. The strength of the DQA process is that it is designed to promote an understanding of how well the data satisfy their intended use by progressing in a logical and efficient manner.

Nevertheless, it should be emphasized that the DQA process cannot *absolutely* prove that one has or has not achieved the DQOs set forth during the planning phase of a study. This situation occurs because a decision maker can never know the *true* value of the item of interest. Data collection only provides the investigators with an *estimate* of this, not its true value. Further, because analytical methods are not perfect, they too can only provide an estimate of the true value of an environmental sample. Because investigators make a decision based on estimated and not true values, they run the risk of making a wrong decision (decision error) about the item of interest.

0.4 INTENDED AUDIENCE

This guidance is written for a broad audience of potential data users, data analysts, and data generators. Data users (such as project managers, risk assessors, or principal investigators who are responsible for making decisions or producing estimates regarding environmental characteristics based on environmental data) should find this guidance useful for understanding and directing the technical work of others who produce and analyze data. Data analysts (such as quality assurance specialists, or any technical professional who is responsible for evaluating the quality of environmental data) should find this guidance to be a convenient compendium of basic assessment tools. Data generators (such as analytical chemists, field sampling specialists, or technical support staff responsible for collecting and analyzing environmental samples and reporting the resulting data values) should find this guidance useful for understanding how their work will be used and for providing a foundation for improving the efficiency and effectiveness of the data generation process.

0.5 ORGANIZATION

This guidance presents background information and statistical tools for performing DQA. Each chapter corresponds to a step in the DQA Process and begins with an overview of the activities to be performed for that step. Following the overviews in Chapters 1, 2, 3, and 4, specific graphical or statistical tools are described and step-by-step procedures are provided along with examples.

0.6 SUPPLEMENTAL SOURCES

Many of the graphical and statistical tools presented in this guidance are also implemented in a user-friendly, personal computer software program called DataQUEST (Data Quality Evaluation Statistical Tools, EPA QA/G-9D). DataQUEST simplifies the implementation of DQA by automating many of the recommended statistical tools. DataQUEST runs on most IBM-compatible personal computers using the DOS operating system; see the DataQUEST User's Guide for complete information on the minimum computer requirements.

The main references in this document are important works having application to environmental sampling and interpretation of data; most of these references are widely available within the scientific and environmental communities. The remaining references are either more detailed original academic articles or are not as readily available to analysts. Two excellent Agency references for analyzing environmental data are *Guidance on the Statistical Analysis of Ground-Water Monitoring Data* (EPA 1992a), a useful compendium of statistical methods and procedures (many of which are incorporated in this document) for the analysis of data generated by EPA's Office of Solid Waste; and *Scout: A Data Analysis Program* (EPA 1993b), a software program for analyzing multivariate data that includes methods for identifying multivariate outliers, graphing the raw data, and displaying the results of principal component analysis.

0.7 SCOPE AND LIMITATIONS

This guidance is intended to be a convenient compendium of practical methods for the environmental scientist and manager. It focuses on measurement data obtained through sampling and analysis of contaminants in environmental media. Statistical nomenclature has been kept to the minimum and there are some areas that will require the input of an environmental statistician for complete analysis. The intent of the document is to assist the non-statistician in the review and analysis of environmental data.

This document represents the first edition of the DQA guidance, which will be followed by annual updates. Readers are encouraged to send their suggestions for improvements and additions to the U.S. EPA Quality Assurance Division. (The address is given in the Foreword.) The annual updates will refine existing sections, present new tools and procedures, and expand the scope of application to additional types of environmental problems.

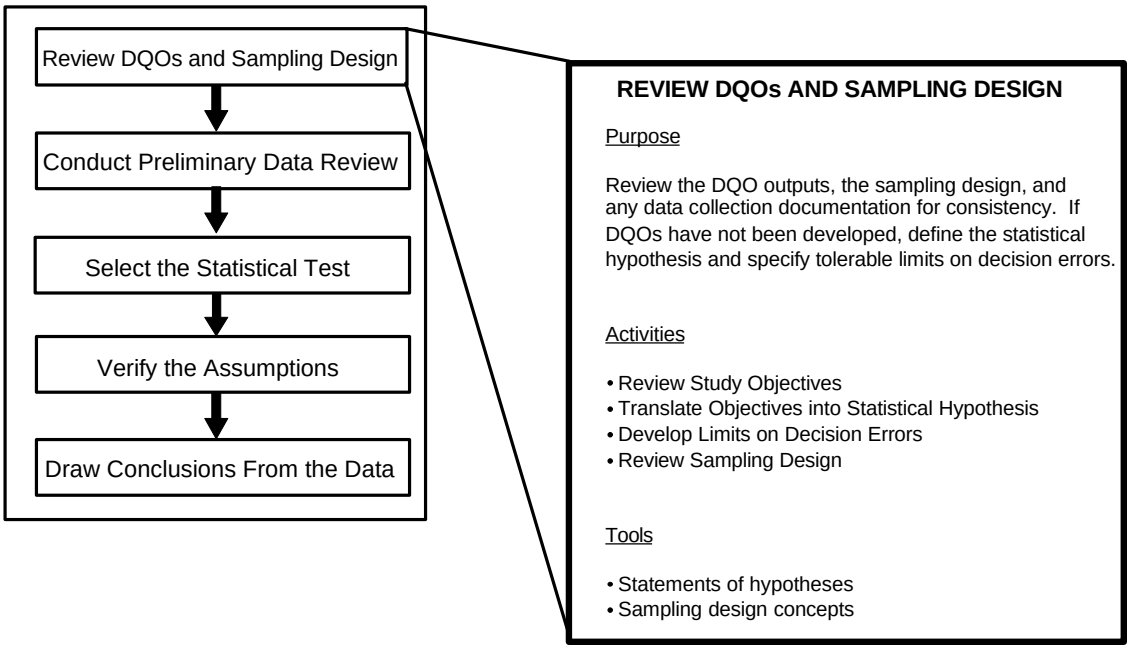
This first edition is intended to cover most of the core topics of DQA for regulatory compliance decisions that involve spatially distributed contamination. Most of the tools will also be applicable to sampling data from hazardous waste sites or facilities under Superfund or RCRA. Many of the tools are generally applicable and useful for other types of problems as well. Future editions of this guidance will address more thoroughly the problems and issues associated with analyzing sampling data from more dynamic processes, such as effluent discharged to waterways and emissions dispersed in ambient air. Future editions will also address other topics, such as analyzing results from designed experiments and other research studies, as well as environmental enforcement investigations.

This guidance is explicitly *not* intended to cover certain topics that are important in some areas of environmental protection. For example, it does not address the important area of survey sampling involving the administration of interviews or questionnaires to people. This document is not intended to substitute for more thorough treatments of fundamental statistical concepts (as found in standard textbooks), nor is it intended to provide a forum for publishing original research (as found in scholarly journals).

CHAPTER 1

STEP 1: REVIEW DQOs AND THE SAMPLING DESIGN

THE DATA QUALITY ASSESSMENT PROCESS



Step 1: Review DQOs and Sampling Design

- Review the objectives of the study.
  - If DQOs have not been developed, review section 1.1.1 and define these objectives.
  - If DQOs were developed, review the outputs from the DQO Process.
- Translate the data user's objectives into a statement of the primary statistical hypothesis.
  - If DQOs have not been developed, review sections 1.1.2 and 1.2, and Table 1.2-1, then develop a statement of the hypothesis based on the data user's objectives.
  - If DQOs were developed, translate them into a statement of the primary hypothesis.
- Translate the data user's objectives into limits on Type I or Type II decision errors.
  - If DQOs have not been developed, review section 1.1.3 and document the data user's tolerable limits on decision errors.
  - If DQOs were developed, confirm the limits on decision errors.
- Review the sampling design and note any special features or potential problems.
  - Review the sampling design for any deviations (sections 1.1.4 and 1.3).

STEP 1: REVIEW DQOs AND THE SAMPLING DESIGN

		Page
1.1	OVERVIEW AND ACTIVITIES	1.1 - 1
1.1.1	Review Study Objectives	1.1 - 1
1.1.2	Translate Objectives into Statistical Hypotheses	1.1 - 2
1.1.3	Develop Limits on Decision Errors	1.1 - 2
1.1.4	Review Sampling Design	1.1 - 3
1.2	DEVELOPING THE STATEMENT OF HYPOTHESES	1.2 - 1
1.3	DESIGNS FOR SAMPLING ENVIRONMENTAL MEDIA	1.3 - 1
1.3.1	Authoritative Sampling	1.3 - 1
1.3.2	Probability Sampling	1.3 - 1
1.3.2.1	Simple Random Sampling	1.3 - 1
1.3.2.2	Sequential Random Sampling	1.3 - 2
1.3.2.3	Systematic Samples	1.3 - 2
1.3.2.4	Stratified Samples	1.3 - 2
1.3.2.5	Compositing Physical Samples	1.3 - 3
1.3.2.6	Other Sampling Designs	1.3 - 3

LIST OF TABLES

Table No.	Page
1.2-1. Commonly Used Statements of Statistical Hypotheses	1.2 - 3

LIST OF BOXES

Box No.	Page
1.1-1: Example Applying the DQO Process Retrospectively	1.1 - 4

Probability Sampling Designs	Section
Simple Random Sampling	1.3.2.1
Sequential Random Sampling	1.3.2.2
Systematic Samples	1.3.2.3
Stratified Samples	1.3.2.4
Compositing Physical Samples	1.3.2.5
Adaptive Sampling	1.3.2.6
Ranked Set Sampling	1.3.2.6

CHAPTER 1
STEP 1: REVIEW DQOs AND THE SAMPLING DESIGN

1.1 OVERVIEW AND ACTIVITIES

The DQA Process begins by reviewing the key outputs from the planning phase of the data life cycle: the Data Quality Objectives (DQOs), the Quality Assurance Project Plan (QAPP), and any associated documents. The DQOs provide the context for understanding the purpose of the data collection effort and establish the qualitative and quantitative criteria for assessing the quality of the data set for the intended use. The sampling design (documented in the QAPP) provides important information about how to interpret the data. By studying the sampling design, the analyst can gain an understanding of the assumptions under which the design was developed, as well as the relationship between these assumptions and the DQOs. By reviewing the methods by which the samples were collected, measured, and reported, the analyst prepares for the preliminary data review and subsequent steps of the DQA Process.

Careful planning improves the representativeness and overall quality of a sampling design, the effectiveness and efficiency with which the sampling and analysis plan is implemented, and the usefulness of subsequent DQA efforts. Given the benefits of planning, the Agency has developed the DQO Process which is a logical, systematic planning procedure based on the scientific method. The DQO Process emphasizes the planning and development of a sampling design to collect the right type, quality, and quantity of data needed to support the decision. Using both the DQO Process and the DQA Process will help to ensure that the decisions are supported by data of adequate quality; the DQO Process does so *prospectively* and the DQA Process does so *retrospectively*.

When DQOs have not been developed during the planning phase of the study, it is necessary to develop statements of the data user's objectives prior to conducting DQA. The primary purpose of stating the data user's objectives prior to analyzing the data is to establish appropriate criteria for evaluating the quality of the data with respect to their intended use. Analysts who are not familiar with the DQO Process should refer to the *Guidance for the Data Quality Objectives Process*, EPA QA/G-4 (1994), a book on statistical decision making using tests of hypothesis, or consult a statistician.

The remainder of this chapter addresses recommended activities for performing this step of DQA and technical considerations that support these activities. The remainder of this section describes the recommended activities, the first three of which will differ depending on whether DQOs have already been developed for the study. Section 1.2 describes how to select the null and alternative hypothesis and section 1.3 presents a brief overview of different types of sampling designs.

1.1.1 Review Study Objectives

In this activity, the objectives of the study are reviewed to provide context for analyzing the data. If a planning process has been implemented before the data are collected, then this step reduces to reviewing the documentation on the study objectives. If no planning process was used, the data user should:

- Develop a concise definition of the problem (DQO Process Step 1) and the decision (DQO Process Step 2) for which the data were collected. This should provide the fundamental reason for collecting the environmental data and identify all potential actions that could result from the data analysis.

- Identify if any essential information is missing (DQO Process Step 3). If so, either collect the missing information before proceeding, or select a different approach to resolving the decision.

- Specify the scale of decision making (any subpopulations of interest) and any boundaries on the study (DQO Process Step 4) based on the sampling design. The scale of decision making is the smallest area or time period to which the decision will apply. The sampling design and implementation may restrict how small or how large this scale of decision making can be.

1.1.2 Translate Objectives into Statistical Hypotheses

In this activity, the data user's objectives are used to develop a precise statement of the primary [1] hypotheses to be tested using environmental data. A statement of the primary statistical hypotheses includes a null hypothesis, which is a "baseline condition" that is presumed to be true in the absence of strong evidence to the contrary, and an alternative hypothesis, which bears the burden of proof. In other words, the baseline condition will be retained unless the alternative condition (the alternative hypothesis) is thought to be true due to the preponderance of evidence. In general, such hypotheses consist of the following elements:

- a population parameter of interest, which describes the feature of the environment that the data user is investigating;

- a numerical value to which the parameter will be compared, such as a regulatory or risk-based threshold or a similar parameter from another place (e.g., comparison to a reference site) or time (e.g., comparison to a prior time); and

- the relation (such as "is equal to" or "is greater than") that specifies precisely how the parameter will be compared to the numerical value.

If DQOs were developed, the statement of hypotheses already should be documented in the outputs of Step 6 of the DQO Process. If DQOs have not been developed, then the analyst should consult with the data user to develop hypotheses that address the data user's concerns. Section 1.2 describes in detail how to develop the statement of hypotheses and includes a list of common encountered hypotheses for environmental decisions. .

1.1.3 Develop Limits on Decision Errors

The goal of this activity is to develop numerical probability limits that express the data user's tolerance for committing false positive (Type I) or false negative (Type II) decision errors as a result of uncertainty in the data. A false positive error occurs when the null hypothesis is rejected when it is true. A false negative decision error occurs when the null hypothesis is not rejected when it is false. If tolerable decision error rates were not established prior to data collection, then the data user should:

- Specify the gray region where the consequences of a false negative decision error are relatively minor (DQO Process Step 6). The gray region is bounded on one side by the threshold value and on the other

---

[1] Throughout this document, the term "primary hypotheses" refers to the statistical hypotheses that correspond to the data user's decision. Other statistical hypotheses can be formulated to formally test the *assumptions* that underlie the specific calculations used to test the primary hypotheses. See Chapter 3 for examples of assumptions underlying primary hypotheses and Chapter 4 for examples of how to test these underlying assumptions.

side by that parameter value where the consequences of making a false negative decision error begin to be significant. Establish this boundary by evaluating the consequences of not rejecting the null hypothesis when it is false and then place the edge of the gray region where these consequences are severe enough to set a limit on the magnitude of this false negative decision error. The gray region is the area between this parameter value and the threshold value.

The width of the gray region represents one important aspect of the decision maker's concern for decision errors. A more narrow gray region implies a desire to detect conclusively the condition when the true parameter value is close to the threshold value ("close" relative to the variability in the data). When the true value of the parameter falls within the gray region, the decision maker may face a high probability of making a false negative decision error, because the data may not provide conclusive evidence for rejecting the null hypothesis, even though it is false (i.e., the data may be too variable to allow the decision maker to recognize that the baseline condition is, in fact, *not* true).

- Specify tolerable limits on the probability of committing false positive and false negative decision errors (DQO Process Step 6) that reflect the decision maker's tolerable limits for making an incorrect decision. Select a possible value of the parameter; then, choose a probability limit based on an evaluation of the seriousness of the potential consequences of making the decision error if the true parameter value is located at that point. At a minimum, the decision maker should specify a false positive decision error limit at the threshold value (α), and a false negative decision error limit at the other edge of the gray region (β).

An example of the gray region and limits on the probability of committing both false positive and false negative decision errors are contained in Box 1.1-1.

If DQOs were developed for the study, the tolerable limits on decision errors will already have been developed. These values can be transferred directly as outputs for this activity. In this case, the action level is the threshold value; the false positive error rate at the action level is the Type I error rate or α; and the false negative error rate at the other bound of the gray region is the Type II error rate or β.

1.1.4 Review Sampling Design

The goal of this activity is to familiarize the analyst with the main features of the sampling design that was used to generate the environmental data. The overall type of sampling design and the manner in which samples were collected or measurements were taken will place conditions and constraints on how the data must be used and interpreted. Section 1.3 provides additional information about several different types of sampling designs that are commonly used in environmental studies.

Review the sampling design documentation with the data user's objectives in mind. Look for design features that support or contradict those objectives. For example, if the data user is interested in making a decision about the mean level of contamination in an effluent stream over time, then composite samples may be an appropriate sampling approach. On the other hand, if the data user is looking for hot spots of contamination at a hazardous waste site, compositing should only be used with caution, to avoid "averaging away" hot spots. Also, look for potential problems in the implementation of the sampling design. For example, verify that each point in space (or time) had an equal probability of being selected for a simple random sampling design. Small deviations from a sampling plan may have minimal effect on the conclusions drawn from the data set. Significant or substantial deviations should be flagged and their potential effect carefully considered throughout the entire DQA.

Box 1.1-1: Example Applying the DQO Process Retrospectively

A waste incineration company was concerned that waste fly ash could contain hazardous levels of cadmium and should be disposed of in a RCRA landfill. As a result, eight composite samples each consisting of eight grab samples were taken from each load of waste. The TCLP leachate from these samples were then analyzed using a method specified in 40 CFR, Pt. 261, App. II. DQOs were not developed for this problem; therefore, study objectives (sections 1.1.1 through 1.1.3) should be developed before the data are analyzed.

1.1.1 Review Study Objectives

- Develop a concise definition of the problem – The problem is defined above.
- Identify if any essential information is missing – It does not appear than any essential information is missing.
- Specify the scale of decision making – Each waste load is sampled separately and decisions need to be made for each load. Therefore, the scale of decision making is an individual load.

1.1.2 Translate Objectives into Statistical Hypotheses

Since composite samples were taken, the parameter of interest is the mean cadmium concentration. The RCRA regulatory standard for cadmium in TCLP leachate is 1.0 mg/L. Therefore, the two hypotheses are "mean cadmium $\geq$ 1.0 mg/L" and "mean cadmium < 1.0 mg/L."

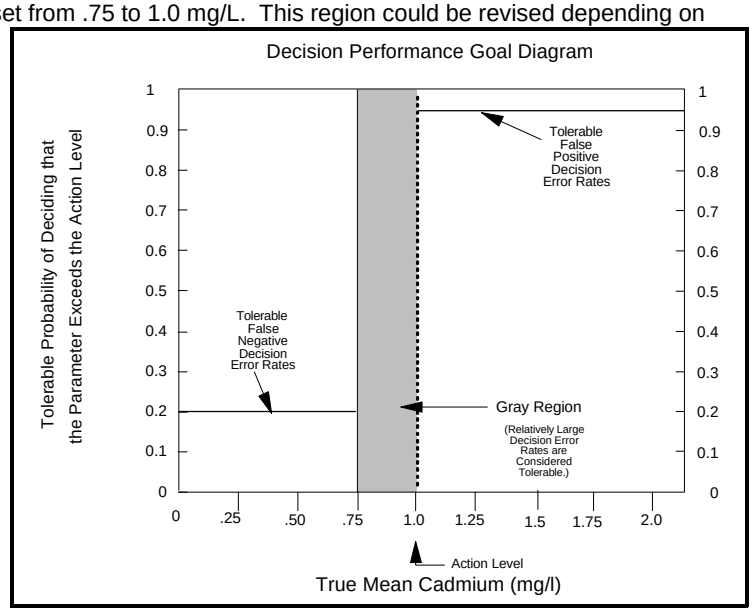
There are two possible decision errors 1) to decide the waste is hazardous ("mean $\geq$ 1.0") when it truly is not ("mean < 1.0"), and 2) to decide the waste is not hazardous ("mean < 1.0") when it truly is ("mean $\geq$ 1.0"). The risk of deciding the fly ash is not hazardous when it truly is hazardous is more severe since potential consequences of this decision error include risk to human health and the environment. Therefore, this error will be labeled the false positive error and the other error will be the false negative error. As a result of this decision, the null hypothesis will be that the waste is hazardous ("mean cadmium $\geq$ 1.0 mg/L") and the alternative hypothesis will be that the waste is not hazardous ("mean cadmium < 1.0 mg/L"). (See section 1.2 for more information on developing the null and alternative hypotheses.)

1.1.3 Develop Limits on Decision Errors

- Specify the gray region – The consequence of a false negative decision error near the action level is unnecessary resource expenditure. The amount of data also influences the width of the gray region. Therefore, for now, a gray region was set from .75 to 1.0 mg/L. This region could be revised depending on the power of the hypothesis test.

- Specify tolerable limits on the probability of committing a decision error – Consequences of a false positive error include risk to human health and environment. Another consequence for the landfill owners is the risk of fines and imprisonment. Therefore, the stringent limit of 0.05 was set on the probability of a false positive decision error. Consequences of a false negative error include unnecessary expenditures so a limit of 0.20 was set on its probability. This error rate could be revised based on the power of the hypothesis test.

The results of this planning process are summarized in the Decision Performance Goal Diagram.



1.2 DEVELOPING THE STATEMENT OF HYPOTHESES

The full statement of the statistical hypotheses has two major parts: the null hypothesis (H_0) and the alternative hypothesis (H_A). In both parts, a population parameter is compared to either a fixed value (for a one-sample test) or another population parameter (for a two-sample test). The population parameter is a quantitative characteristic of the population that the data user wants to estimate using the data. In other words, the parameter describes that feature of the population that the data user will evaluate when making the decision. Examples of parameters are the population mean and median.

If the data user is interested in drawing inferences about only one population, then the null and alternative hypotheses will be stated in terms that relate the true value of the parameter to some fixed threshold value. A common example of this one-sample problem in environmental studies is when pollutant levels in an effluent stream are compared to a regulatory limit. If the data user is interested in comparing two populations, then the null and alternative hypotheses will be stated in terms that compare the true value of one population parameter to the corresponding true parameter value of the other population. A common example of this two-sample problem in environmental studies is when a potentially contaminated waste site is being compared to a reference area using samples collected from the respective areas. In this situation, the hypotheses often will be stated in terms of the difference between the two parameters.

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

An alternative method for defining the null hypothesis is based on historical information. If a large amount of information exists suggesting that one hypothesis is extremely likely, then this hypothesis should be defined as the alternative hypothesis. In this case, a large amount of data may not be necessary to provide overwhelming evidence that the other (null) hypothesis is false. For example, if the waste from an incinerator was previously hazardous and the waste process has not changed, it may be more cost-effective to define the alternative hypothesis as "the waste is hazardous" and the null hypothesis as "the waste is not hazardous."

Consider a data user who wants to know whether the true mean concentration (μ) of atrazine in ground water at a hazardous waste site is greater than a fixed threshold value C. If the data user presumes from prior information that the true mean concentration is at least C due possibly to some contamination incident, then the data must provide compelling evidence to reject that presumption, and the hypotheses can be stated as follows:

Narrative Statement of Hypotheses	Statement of Hypotheses Using Standard Notation
Null Hypothesis (Baseline Condition): The true mean concentration of atrazine in ground water is greater than or equal to the threshold value C; versus	H_0: $\mu \geq C$; versus
Alternative Hypothesis: The true mean concentration of atrazine in ground water is less than the threshold value C.	H_A: $\mu < C$

On the other hand, if the data user presumes from prior information that the true mean concentration is less than C due possibly to the fact that the ground water has not been contaminated in the past, then the data must provide compelling evidence to reject that presumption, and the hypotheses can be stated as follows:

Narrative Statement of Hypotheses	Statement of Hypotheses Using Standard Notation
Null Hypothesis (Baseline Condition): The true mean concentration of atrazine in ground water is less than or equal to the threshold value C; versus	H_0: $\mu \leq C$; versus
Alternative Hypothesis: The true mean concentration of atrazine in ground water is greater than the threshold value C.	H_A: $\mu > C$

In stating the primary hypotheses, it is convenient to use standard statistical notation, as shown throughout this document. However, the logic underlying the hypothesis always corresponds to the decision of interest to the data user.

Table 1.2-1 summarizes some common types of environmental decisions and the corresponding hypotheses. In Table 1.2-1, the parameter is denoted using the symbol "Θ," and the difference between two parameters is denoted using "$\Theta_1 - \Theta_2$" where Θ_1 represents the parameter of the first population and Θ_2 represents the parameter of the second population. The use of "Θ" is to avoid using the terms "population mean" or "population median" repeatedly because the structure of the hypothesis test remains the same regardless of the population parameter. The fixed threshold value is denoted "C," and the difference between two parameters is denoted "δ_0" (it is common to see the null hypothesis defined such that $\delta_0 = 0$). If the data user's problem does not fall into one of the categories described in Table 1.2-1, the problem and associated hypotheses may be of a more complicated form and a statistician should be consulted.

For the first of the six decision problems in Table 1.2-1, only estimates of Θ that exceed C can cast doubt on the null hypothesis. This is called a one-tailed hypothesis test, because only parameter estimates on one side of the threshold value can lead to rejection of the null hypothesis. The second, fourth, and fifth rows of Table 1.2-1 are also examples of one-tailed hypothesis tests. The third and sixth rows of Table 1.2-1 are examples of two-tailed tests, because estimates falling both below and above the null-hypothesis parameter value can lead to rejection of the null hypothesis. Most hypotheses connected with environmental monitoring are one-tailed because high pollutant levels can harm humans or ecosystems.

Table 1.2-1. Commonly Used Statements of Statistical Hypotheses

Type of Decision	Null Hypothesis	Alternative Hypothesis
Compare environmental conditions to a fixed threshold value, such as a regulatory standard or acceptable risk level; presume that the true condition is less than the threshold value.	H_0: $\Theta \leq C$	H_A: $\Theta > C$
Compare environmental conditions to a fixed threshold value; presume that the true condition is greater than the threshold value.	H_0: $\Theta \geq C$	H_A: $\Theta < C$
Compare environmental conditions to a fixed threshold value; presume that the true condition is equal to the threshold value and the data user is concerned whenever conditions vary significantly from this value.	H_0: $\Theta = C$	H_A: $\Theta \neq C$
Compare environmental conditions associated with two different populations to a fixed threshold value (δ_0) such as a regulatory standard or acceptable risk level; presume that the true condition is less than the threshold value. If it is presumed that conditions associated with the two populations are the same, the threshold value is 0.	H_0: $\Theta_1 - \Theta_2 \leq \delta_0$ (H_0: $\Theta_1 - \Theta_2 \leq 0$)	H_A: $\Theta_1 - \Theta_2 > \delta_0$ (H_A: $\Theta_1 - \Theta_2 > 0$)
Compare environmental conditions associated with two different populations to a fixed threshold value (δ_0) such as a regulatory standard or acceptable risk level; presume that the true condition is greater than the threshold value. If it is presumed that conditions associated with the two populations are the same, the threshold value is 0.	H_0: $\Theta_1 - \Theta_2 \geq \delta_0$ (H_0: $\Theta_1 - \Theta_2 \geq 0$)	H_A: $\Theta_1 - \Theta_2 < \delta_0$ (H_A: $\Theta_1 - \Theta_2 < 0$)
Compare environmental conditions associated with two different populations to a fixed threshold value (δ_0) such as a regulatory standard or acceptable risk level; presume that the true condition is equal to the threshold value. If it is presumed that conditions associated with the two populations are the same, the threshold value is 0.	H_0: $\Theta_1 - \Theta_2 = \delta_0$ (H_0: $\Theta_1 - \Theta_2 = 0$)	H_A: $\Theta_1 - \Theta_2 \neq \delta_0$ (H_A: $\Theta_1 - \Theta_2 \neq 0$)

1.3 DESIGNS FOR SAMPLING ENVIRONMENTAL MEDIA

Sampling designs provide the basis for how a set of samples may be analyzed. Different sampling designs require different analysis techniques and different assessment procedures. There are two primary types of sampling designs: authoritative (judgment) sampling and probability sampling. This section describes some of the most common sampling designs.

1.3.1 Authoritative Sampling

With authoritative (judgment) sampling, an expert having knowledge of the site (or process) designates where and when samples are to be taken. This type of sampling should only be considered when the objectives of the investigation are not of a statistical nature, for example, when the objective of a study is to identify specific locations of leaks, or when the study is focused solely on the sampling locations themselves. Generally, conclusions drawn from authoritative samples apply only to the individual samples and aggregation may result in severe bias and lead to highly erroneous conclusions. Judgmental sampling also precludes the use of the sample for any purpose other than the original one. Thus if the data may be used in further studies (e.g., for an estimate of variability in a later study), a probabilistic design should be used.

When the study objectives involve estimation or decision making, some form of probability sampling is required. As described below, this does not preclude use of the expert's knowledge of the site or process in designing a probability-based sampling plan; however, valid statistical inferences require that the plan incorporate some form of randomization in choosing the sampling locations or sampling times. For example, to determine maximum SO_2 emission from a boiler, the sampling plan would reasonably focus, or put most of the weight on, periods of maximum or near-maximum boiler operation. Similarly, if a residential lot is being evaluated for contamination, then the sampling plan can take into consideration prior knowledge of contaminated areas, by weighting such areas more heavily in the sample selection and data analysis.

1.3.2 Probability Sampling

Probability samples are samples in which every member of the target population (i.e., every potential sampling unit) has a known probability of being included in the sample. Probability samples can be of various types, but in some way, they all make use of randomization, which allows valid probability statements to be made about the quality of estimates or hypothesis tests that are derived from the resultant data.

One common misconception of probability sampling procedures is that these procedures preclude the use of important prior information. Indeed, just the opposite is true. An efficient sampling design is one that uses all available prior information to stratify the region and set appropriate probabilities of selection. Another common misconception is that using a probability sampling design means allowing the possibility that the sample points will not be distributed appropriately across the region. However, if there is no prior information regarding the areas most likely to be contaminated, a grid sampling scheme (a type of stratified design) is usually recommended to ensure that the sampling points are dispersed across the region.

1.3.2.1 Simple Random Sampling

The simplest type of probability sample is the simple random sample where every possible sampling unit in the target population has an equal chance of being selected. Simple random samples, like the other samples, can be either samples in time and/or space and are often appropriate at an early stage of an

investigation in which little is known about systematic variation within the site or process. All of the sampling units should have equal volume or mass, and ideally be of the same shape if applicable. With a simple random sample, the term "random" should not be interpreted to mean haphazard; rather, it has the explicit meaning of equiprobable selection. Simple random samples are generally developed through use of a random number table or through computer generation of pseudo-random numbers.

1.3.2.2 Sequential Random Sampling

Usually, simple random samples have a fixed sample size, but some alternative approaches are available, such as sequential random sampling, where the sample sizes are not fixed *a priori*. Rather, a statistical test is performed after each specimen's analysis (or after some minimum number have been analyzed). This strategy could be applicable when sampling and/or analysis is quite expensive, when information concerning sampling and/or measurement variability is lacking, when the characteristics of interest are stable over the time frame of the sampling effort, or when the objective of the sampling effort is to test a single specific hypothesis.

1.3.2.3 Systematic Samples

In the case of spatial sampling, systematic sampling involves establishing a two-dimensional (or in some cases a three-dimensional) spatial grid and selecting a random starting location within one of the cells. Sampling points in the other cells are located in a deterministic way relative to that starting point. In addition, the orientation of the grid is sometimes chosen randomly and various types of systematic samples are possible. For example, points may be arranged in a pattern of squares (rectangular grid sampling) or a pattern of equilateral triangles (triangular grid sampling). The result of either approach is a simple pattern of equally spaced points at which sampling is to be performed.

Systematic sampling designs have several advantages over random sampling and some of the other types of probability sampling. They are generally easier to implement, for example. They are also preferred when one of the objectives is to locate "hot spots" within a site or otherwise map the pattern of concentrations over a site. On the other hand, they should be used with caution whenever there is a possibility of some type of cyclical pattern in the waste site or process. Such a situation, combined with the uniform pattern of sampling points, could very readily lead to biased results.

1.3.2.4 Stratified Samples

Another type of probability sample is the stratified random sample, in which the site or process is divided into two or more nonoverlapping strata, sampling units are defined for each stratum, and separate simple random samples are employed to select the units in each stratum. (If a systematic sample were employed within each stratum, then the design would be referred to as a stratified systematic sample.) Strata should be defined so that physical samples within a stratum are more similar to each other than to samples from other strata. If so, a stratified random sample should result in more precise estimates of the overall population parameter than those that would be obtained from a simple random sample with the same number of sampling units.

Stratification is an accepted way to incorporate prior knowledge and professional judgment into a probabilistic sampling design. Generally, units that are "alike" or anticipated to be "alike" are placed together in the same stratum. Units that are contiguous in space (e.g., similar depths) or time are often grouped together into the same stratum, but characteristics other than spatial or temporal proximity can also

be employed. Media, terrain characteristics, concentration levels, previous cleanup attempts, and confounding contaminants can also be used as the basis for creating strata.

Advantages of stratified samples over random samples include their ability to ensure more uniform coverage of the entire target population and, as noted above, their potential for achieving greater precision in certain estimation problems. Even when imperfect information is used to form strata, the stratified random sample will generally be more cost-effective than a simple random sample. A stratified design can also be useful when there is interest in estimating or testing characteristics for subsets of the target population. Because different sampling rates can be used in different strata, one can oversample in strata containing those subareas of particular interest to ensure that they are represented in the sample. In general, statistical calculations for data generated via stratified samples are more complex than for random samples, and certain types of tests, for example, cannot be performed when stratified samples are employed. Therefore, a statistician should be consulted when stratified sampling is used.

1.3.2.5 Compositing Physical Samples

When analysis costs are large relative to sampling costs, cost-effective plans can sometimes be achieved by compositing physical samples or specimens prior to analysis, assuming that there are no safety hazards or potential biases (for example, the loss of volatile organic compounds from a matrix) associated with such compositing. For the same total cost, compositing in this situation would allow a larger number of sampling units to be selected than would be the case if compositing were not used. Composite samples reflect a physical rather than a mathematical mechanism for averaging. Therefore, compositing should generally be avoided if population parameters other than a mean are of interest (e.g., percentiles or standard deviations).

Composite sampling is also useful when the analyses of composited samples are to be used in a two-staged approach in which the composite-sample analyses are used solely as a screening mechanism to identify if additional, separate analyses need to be performed. This situation might occur during an early stage of a study that seeks to locate those areas that deserve increased attention due to potentially high levels of one or more contaminants.

1.3.2.6 Other Sampling Designs

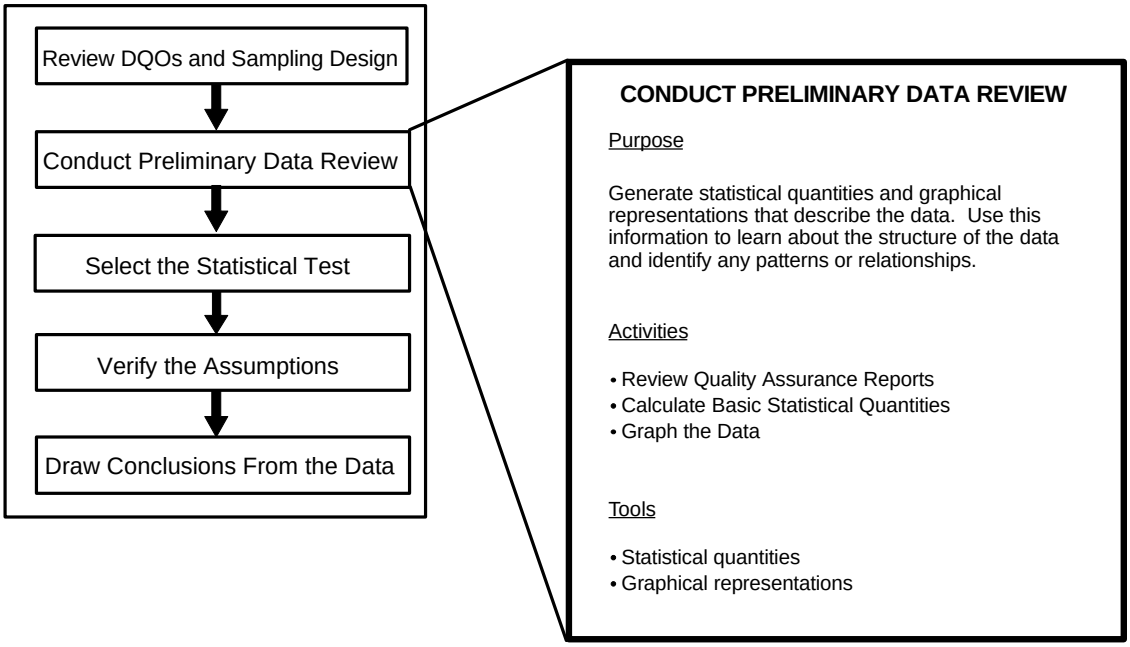
Adaptive sampling involves taking a sample and using the resulting information to design the next stage of sampling. The process may continue through several additional rounds of sampling and analysis. A common application of adaptive sampling to environmental problems involves subdividing the region of interest into smaller units, taking a probability sample of these units, then sampling all units that border on any unit with a concentration level greater than some specified level C. This process is continued until all newly sampled units are below C. The field of adaptive sampling is currently undergoing active development and can be expected to have a significant impact on environmental sampling.

Ranked set sampling (RSS) uses the availability of an inexpensive surrogate measurement when it is correlated with the more expensive measurement of interest. The method exploits this correlation to obtain a sample which is more representative of the population that would be obtained by random sampling, thereby leading to more precise estimates of population parameters than what would be obtained by random sampling. RSS consists of creating n groups, each of size n (for a total of n^2 initial samples), then ranking the surrogate from largest to smallest within each group. One sample from each group is then selected according to a specified procedure and these n samples are analyzed for the more expensive measurement of interest.

CHAPTER 2

STEP 2: CONDUCT A PRELIMINARY DATA REVIEW

THE DATA QUALITY ASSESSMENT PROCESS



Step 2: Conduct a Preliminary Data Review

- Review quality assurance reports.
  - Look for problems or anomalies in the implementation of the sample collection and analysis procedures.
  - Examine QC data for information to verify assumptions underlying the Data Quality Objectives, the Sampling and Analysis Plan, and the Quality Assurance Project Plans.
- Calculate the statistical quantities.
  - Consider calculating appropriate percentiles (section 2.2.1)
  - Select measures of central tendency (section 2.2.2) and dispersion (section 2.2.3).
  - If the data involve two variables, calculate the correlation coefficient (section 2.2.4).
- Display the data using graphical representations.
  - Select graphical representations (section 2.4) that illuminate the structure of the data set and highlight assumptions underlying the Data Quality Objectives, the Sampling and Analysis Plan, and the Quality Assurance Project Plans.
  - Use a variety of graphical representations that examine different features of the set.

STEP 2: CONDUCT A PRELIMINARY DATA REVIEW

Statistical Quantities	Section	Directions	Example
Coefficient of Variation	2.2.3	Box 2.2-4	Box 2.2-5
Correlation Coefficient	2.2.4	Box 2.2-6	Box 2.2-6
Interquartile Range	2.2.3	Box 2.2-4	Box 2.2-5
Mean	2.2.2	Box 2.2-2	Box 2.2-3
Median	2.2.2	Box 2.2-2	Box 2.2-3
Mode	2.2.2	Box 2.2-2	Box 2.2-3
Percentiles/Quantiles	2.2.1	Box 2.2-1	Box 2.2-1
Range	2.2.3	Box 2.2-4	Box 2.2-5
Standard Deviation	2.2.3	Box 2.2-4	Box 2.2-5
Variance	2.2.3	Box 2.2-4	Box 2.2-5

Graphical Representations	Section	Figure	Directions	Example
Box and Whisker Plot	2.3.3	Figure 2.3-3	Box 2.3-5	Box 2.3-6
Coded Scatter Plot.	2.3.7.3	Figure 2.3-9		
Contour Plots	2.3.9.3			
Autocorrelation Function	2.3.8.2	Figure 2.3-13	Box 2.3-16	Box 2.3-17
Empirical Quantile-Quantile Plot	2.3.7.4	Box 2.3-14	Box 2.3-14	Box 2.3-14
Frequency Plots	2.3.1	Figure 2.3-1	Box 2.3-1	Box 2.3-2
h-Scatterplot	2.3.9.3			
Histogram	2.3.1	Figure 2.3-2	Box 2.3-1	Box 2.3-2
Normal Probability Plot	2.3.6	Box 2.3-12	Box 2.3-11	Box 2.3-12
Parallel Coordinate Plot	2.3.7.3	Figure 2.3-10		
Posting Plots	2.3.9.1	Figure 2.3-14	Box 2.3-18	Box 2.3-18
Quantile Plot	2.3.5	Figure 2.3-5	Box 2.3-9	Box 2.3-10
Ranked Data Plot	2.3.4	Figure 2.3-4	Box 2.3-7	Box 2.3-8
Scatter Plot	2.3.7.2	Figure 2.3-8	Box 2.3-13	Box 2.3-13
Scatter Plot Matrix	2.3.7.3	Figure 2.3-11		
Stem-and-leaf Plot	2.3.2	Box 2.3-4	Box 2.3-3	Box 2.3-4
Symbol Plots	2.3.9.2	Figure 2.3-15	Box 2.3-18	Box 2.3-18
Time Plot	2.3.8.1	Figure 2.3-12	Box 2.3-15	Box 2.3-15

CHAPTER 2
STEP 2: CONDUCT A PRELIMINARY DATA REVIEW

2.1 OVERVIEW AND ACTIVITIES

In this step of the DQA Process, the analyst conducts a preliminary evaluation of the data set, calculates some basic statistical quantities, and examines the data using graphical representations. A preliminary data review should be performed whenever data are used, regardless of whether they are used to support a decision, estimate a population parameter, or answer exploratory research questions. By reviewing the data both numerically and graphically, one can learn the "structure" of the data and thereby identify appropriate approaches and limitations for using the data. The DQA software DataQUEST (G-9D, 1996) will perform all of these functions as well as more sophisticated statistical tests.

There are two main elements of preliminary data review: (1) basic statistical quantities (summary statistics); and (2) graphical representations of the data. Statistical quantities are functions of the data that numerically describe the data set. Examples include a mean, median, percentile, range, and standard deviation. They can be used to provide a mental picture of the data and are useful for making inferences concerning the population from which the data were drawn. Graphical representations are used to identify patterns and relationships within the data, confirm or disprove hypotheses, and identify potential problems. For example, a normal probability plot may allow an analyst to quickly discard an assumption of normality and may identify potential outliers.

The preliminary data review step is designed to make the analyst familiar with the data. The review should identify anomalies that could indicate unexpected events that may influence the analysis of the data. The analyst may know what to look for based on the anticipated use of the data documented in the Data Quality Objectives Process, the Quality Assurance Project Plan, and any associated documents. The results of the review are then used to select a procedure for testing a statistical hypotheses to support the data user's decision.

2.1.1 Review Quality Assurance Reports

The first activity in conducting a preliminary data review is to review any relevant quality assurance (QA) reports that describe the data collection and reporting process as it actually was implemented. These QA reports provide valuable information about potential problems or anomalies in the data set. Specific items that may be helpful include:

- Data validation reports that document the sample collection, handling, analysis, data reduction, and reporting procedures used;

- Quality control reports from laboratories or field stations that document measurement system performance, including data from check samples, split samples, spiked samples, or any other internal QC measures; and

- Technical systems reviews, performance evaluation audits, and audits of data quality, including data from performance evaluation samples.

When reviewing QA reports, particular attention should be paid to information that can be used to check assumptions made in the Data Quality Objectives Process. Of great importance are apparent anomalies in recorded data, missing values, deviations from standard operating procedures, and the use of nonstandard data collection methodologies.

2.1.2 Calculate Basic Statistical Quantities

The goal of this activity is to summarize some basic quantitative characteristics of the data set using common statistical quantities. Some statistical quantities that are useful to the analyst include: number of observations; measures of central tendency, such as a mean, median, or mode; measures of dispersion, such as range, variance, standard deviation, coefficient of variation, or interquartile range; measures of relative standing, such as percentiles; measures of distribution symmetry or shape; and measures of association between two or more variables, such as correlation. These measures can then be used for description, communication, and to test hypothesis regarding the population from which the data were drawn. Section 2.2 provides detailed descriptions and examples of these statistical quantities.

The sample design may influence how the statistical quantities are computed. The formulas given in this chapter are for simple random sampling, simple random sampling with composite samples, and randomized systematic sampling. If a more complex design is used, such as a stratified design, then the formulas may need to be adjusted.

2.1.3 Graph the Data

The goal of this step is to identify patterns and trends in the data that might go unnoticed using purely numerical methods. Graphs can be used to identify these patterns and trends, to quickly confirm or disprove hypotheses, to discover new phenomena, to identify potential problems, and to suggest corrective measures. In addition, some graphical representations can be used to record and store data compactly or to convey information to others. Graphical representations include displays of individual data points, statistical quantities, temporal data, spatial data, and two or more variables. Since no single graphical representation will provide a complete picture of the data set, the analyst should choose different graphical techniques to illuminate different features of the data. Section 2.3 provides descriptions and examples of common graphical representations.

At a minimum, the analyst should choose a graphical representation of the individual data points and a graphical representation of the statistical quantities. If the data set has a spatial or temporal component, select graphical representations specific to temporal or spatial data in addition to those that do not. If the data set consists of more than one variable, treat each variable individually before developing graphical representations for the multiple variables. If the sampling plan or suggested analysis methods rely on any critical assumptions, consider whether a particular type of graph might shed light on the validity of that assumption. For example, if a small-sample study is strongly dependent on the assumption of normality, then a normal probability plot would be useful (section 2.3.6).

The sampling design may influence what data may be included in each representation. Usually, the graphical representations should be applied to each complete unit of randomization separately or each unit of randomization should be represented with a different symbol. For example, the analyst could generate box plots for each stratum instead of generating one box plot that includes the data from all the strata.

2.2 STATISTICAL QUANTITIES

2.2.1 Measures of Relative Standing

Sometimes the analyst is interested in knowing the relative position of one of several observations in relation to all of the observations. Percentiles are one such measure of relative standing that may also be useful for summarizing data. A percentile is the data value that is greater than or equal to a given percentage of the data values. Stated in mathematical terms, the pth percentile is the data value that is greater than or equal to p% of the data values and is less than or equal to (1-p)% of the data values. Therefore, if 'x' is the pth percentile, then p% of the values in the data set are less than or equal to x, and (100-p)% of the values are greater than or equal to x. A sample percentile may fall between a pair of observations. For example, the 75th percentile of a data set of 10 observations is not uniquely defined. Therefore, there are several methods for computing sample percentiles, the most common of which is described in Box 2.2-1.

Important percentiles usually reviewed are the quartiles of the data, the 25th, 50th, and 75th percentiles. The 50th percentile is also called the sample median (section 2.2.2), and the 25th and 75th percentile are used to estimate the dispersion of a data set (section 2.2.3). Also important for environmental data are the 90th, 95th, and 99th percentile where a decision maker would like to be sure that 90%, 95%, or 99% of the contamination levels are below a fixed risk level.

Box 2.2-1: Directions for Calculating the Measure of Relative Standing (Percentiles) with an Example

Let $X_1, X_2, ..., X_n$ represent the n data points. To compute the pth percentile, y(p), first list the data from smallest to largest and label these points $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ (so that $X_{(1)}$ is the smallest, $X_{(2)}$ is the second smallest, and $X_{(n)}$ is the largest). Let t = p/100, and multiply the sample size n by t. Divide the result into the integer part and the fractional part, i.e., let nt = j + g where j is the integer part and g is the fraction part. Then the pth percentile, y(p), is calculated by:

If g = 0, $y(p) = (X_{(j)} + X_{(j+1)})/2$

otherwise, $y(p) = X_{(j+1)}$

Example: The 90th and 95th percentile will be computed for the following 10 data points (ordered from smallest to largest) : 4, 4, 4, 5, 5, 6, 7, 7, 8, and 10 ppb.

For the 95th percentile, t = p/100 = 95/100= .95 and nt = (10)(.95) = 9.5 = 9 + .5. Therefore, j = 9 and g = .5. Because $g = .5 \neq 0$, $y(95) = X_{(j+1)} = X_{(9+1)} = X_{(10)}$ = 10 ppm. Therefore, 10 ppm is the 95th percentile of the above data.

For the 90th percentile, t = p/100 = 90/100 = .9 and nt = (10)(.9) = 9. Therefore j = 9 and g = 0. Since g = 0, $y(90) = (X_{(9)} + X_{(10)}) / 2 = (8 + 10) / 2 = 9$ ppm.

A quantile is similar in concept to a percentile; however, a percentile represents a percentage whereas a quantile represents a fraction. If 'x'is the pth percentile, then at least p% of the values in the data set lie at or below x, and at least (100-p)% of the values lie at or above x, whereas if x is the p/100 quantile of the data, then the fraction p/100 of the data values lie at or below x and the fraction (1-p)/100 of the data values lie at or above x. For example, the .95 quantile has the property that .95 of the observations lie at or below x and .05 of the data lie at or above x. For the example in Box 2.2-1, 9 ppm would be the .95 quantile and 10 ppm would be the .99 quantile of the data.

2.2.2 Measures of Central Tendency

Measures of central tendency characterize the center of a sample of data points. The three most common estimates are the mean, median, and the mode. Directions for calculating these quantities are contained in Box 2.2-2; examples are provided in Box 2.2-3.

The most commonly used measure of the center of a sample is the sample mean, denoted by $\bar{x}$. This estimate of the center of a sample can be thought of as the "center of gravity" of the sample. The sample mean is an arithmetic average for simple sampling designs; however, for complex sampling designs, such as stratification, the sample mean is a weighted arithmetic average. The sample mean is influenced by extreme values (large or small) and nondetects (see section 4.7).

The sample median ($\tilde{x}$) is the second most popular measure of the center of the data. This value falls directly in the middle of the data when the measurements are ranked in order from smallest to largest. This means that ½ of the data are smaller than the sample median and ½ of the data are larger than the sample median. The median is another name for the 50 th percentile (section 2.2.1). The median is not influenced by extreme values and can easily be used in the case of censored data (nondetects).

The third method of measuring the center of the data is the mode. The sample mode is the value of the sample that occurs with the greatest frequency. Since this value may not always exist, or if it does it may not be unique, this value is the least commonly used. However, the mode is useful for qualitative data.

2.2.3 Measures of Dispersion

Measures of central tendency are more meaningful if accompanied by information on how the data spread out from the center. Measures of dispersion in a data set include the range, variance, sample standard deviation, coefficient of variation, and the interquartile range. Directions for computing these measures are given in Box 2.2-4; examples are given in Box 2.2-5.

The easiest measure of dispersion to compute is the sample range. For small samples, the range is easy to interpret and may adequately represent the dispersion of the data. For large samples, the range is not very informative because it only considers (and therefore is greatly influenced) by extreme values.

The sample variance measures the dispersion from the mean of a data set. A large sample variance implies that there is a large spread among the data so that the data are not clustered around the mean. A small sample variance implies that there is little spread among the data so that most of the data are near the mean. The sample variance is affected by extreme values and by a large number of nondetects. The sample standard deviation is the square root of the sample variance and has the same unit of measure as the data.

The coefficient of variation (CV) is a unitless measure that allows the comparison of dispersion across several sets of data. The CV is often used in environmental applications because variability (expressed as a standard deviation) is often proportional to the mean.

When extreme values are present, the interquartile range may be more representative of the dispersion of the data than the standard deviation. This statistical quantity does not depend on extreme values and is therefore useful when the data include a large number of nondetects.

Box 2.2-2: Directions for Calculating the Measures of Central Tendency

Let X_1, X_2, ..., X_n represent the n data points.

Sample Mean: The sample mean $\bar{x}$ is the sum of all the data points divided by the total number of data points (n):

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Sample Median: The sample median ($\tilde{X}$) is the center of the data when the measurements are ranked in order from smallest to largest. To compute the sample median, list the data from smallest to largest and label these points $X_{(1)}$, $X_{(2)}$, . . ., $X_{(n)}$ (so that $X_{(1)}$ is the smallest, $X_{(2)}$ is the second smallest, and $X_{(n)}$ is the largest).

If the number of data points is odd, then $\tilde{X} = X_{([n+1]/2)}$

If the number of data points is even, then $\tilde{X} = \dfrac{X_{(n/2)} + X_{([n/2]+1)}}{2}$

Sample Mode: The mode is the value of the sample that occurs with the greatest frequency. The mode may not exist, or if it does, it may not be unique. To find the mode, count the number of times each value occurs. The sample mode is the value that occurs most frequently.

Box 2.2-3: Example Calculations of the Measures of Central Tendency

Using the directions in Box 2.2-2 and the following 10 data points (in ppm): 4, 5, 6, 7, 4, 10, 4, 5, 7, and 8, the following is an example of computing the sample mean, median, and mode.

Sample mean:

$$\bar{X} = \frac{4 + 5 + 6 + 7 + 4 + 10 + 4 + 5 + 7 + 8}{10} = \frac{60}{10} = 6 \ ppm$$

Therefore, the sample mean is 6 ppm.

Sample median: The ordered data are: 4, 4, 4, 5, 5, 6, 7, 7, 8, and 10. Since n=10 is even, the sample median is

$$\tilde{X} = \frac{X_{(10/2)} + X_{([10/2]+1)}}{2} = \frac{X_{(5)} + X_{(6)}}{2} = \frac{5 + 6}{2} = 5.5 \ ppm$$

Thus, the sample median is 5.5 ppm.

Sample mode: Computing the number of times each value occurs yields:

4 appears 3 times; 5 appears 2 times; 6 appears 1 time; 7 appears 2 times; 8 appears 1 time; and 10 appears 1 time.

Because the value of 4 ppm appears the most times, it is the mode of this data set.

Box 2.2-4: Directions for Calculating the Measures of Dispersion

Let $X_1, X_2, ..., X_n$ represent the n data points.

Sample Range: The sample range (R) is the difference between the largest value and the smallest value of the sample, i.e., R = maximum - minimum.

Sample Variance: To compute the sample variance (s^2), compute:

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2}{n-1}$$

Sample Standard Deviation: The sample standard deviation (s) is the square root of the sample variance, i.e.,

$$s = \sqrt{s^2}$$

Coefficient of Variation: The coefficient of variation (CV) is the standard deviation divided by the sample mean (section 2.2.2), i.e., CV = s/$\bar{x}$. The CV is often expressed as a percentage.

Interquartile Range: Use the directions in section 2.2.1 to compute the 25th and 75th percentiles of the data (y(25) and y(75) respectively). The interquartile range (IQR) is the difference between these values, i.e., IQR = y(75) - y(25).

Box 2.2-5: Example Calculations of the Measures of Dispersion

In this box, the directions in Box 2.2-4 and the following 10 data points (in ppm): 4, 5, 6, 7, 4, 10, 4, 5, 7, and 8, are used to calculate the measures of dispersion. From Box 2.2-2, $\bar{x}$ = 6 ppm.

Sample Range: R = maximum - minimum = 10 - 4 = 6 ppm

Sample Variance:

$$s^2 = \frac{[4^2+5^2+...+7^2+8^2] - \frac{(4+5+...+7+8)^2}{10}}{10-1} = \frac{396 - \frac{(60)^2}{10}}{9} = 4\ ppm^2$$

Sample Standard Deviation: $s = \sqrt{s^2} = \sqrt{4} = 2\ ppm$

Coefficient of Variation: $CV = s / \bar{X} = 2ppm / 6ppm = \frac{1}{3} = 33\%$

Interquartile Range: Using the directions in section 2.2.1 to compute the 25th and 75th percentiles of the data (y(25) and y(75) respectively): y(25) = $X_{(2+1)} = X_{(3)}$ = 4 ppm and y(75) = $X_{(7+1)} = X_{(8)}$ = 7 ppm. The interquartile range (IQR) is the difference between these values: IQR = y(75) - y(25) = 7 - 4 = 3 ppm

2.2.4 Measures of Association

Data often include measurements of several characteristics (variables) for each sample point and there may be interest in knowing the relationship or level of association between two or more of these variables. One of the most common measures of association is the correlation coefficient. Directions and an example for calculating a correlation coefficient are contained in Box 2.2-6.

The correlation coefficient measures the linear relationship between two variables. A linear association implies that as one variable increases so does the other linearly, or as one variable decreases the other increases linearly. Values of the correlation coefficient close to +1 (positive correlation) imply that as one variable increases so does the other, the reverse holds for values close to -1. A value of +1 implies a perfect positive linear correlation, i.e., all the data pairs lie on a straight line with a positive slope. A value of -1 implies perfect negative linear correlation. Values close to 0 imply little correlation between the variables.

The correlation coefficient does not imply cause and effect. The analyst may say that the correlation between two variables is high and the relationship is strong, but may not say that one variable causes the other variable to increase or decrease without further evidence and strong statistical controls. The correlation coefficient does not detect nonlinear relationships so it should be used only in conjunction with a scatter plot (section 2.3.7.2). A scatter plot can be used to determine if the correlation coefficient is meaningful or if some measure of nonlinear relationships should be used. The correlation coefficient can be significantly changed by extreme values so a scatter plot should be used first to identify such values.

Box 2.2-6: Directions for Calculating the Correlation Coefficient with an Example

Let $X_1, X_2, ..., X_n$ represent one variable of the n data points and let $Y_1, Y_2, ..., Y_n$ represent a second variable of the n data points. The Pearson correlation coefficient, r, between X and Y is computed by:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\left[[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}] \; [\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}] \right]^{1/2}}$$

<u>Example</u>: Consider the following data set (in ppb): Sample 1 — arsenic (X) = 4.0, lead (Y) = 8.0; Sample 2 - arsenic = 3.0, lead = 7.0; Sample 3 - arsenic = 2.0, lead = 7.0; and Sample 4 - arsenic = 1.0, lead = 6.0.

$$\sum_{i=1}^{n} X_i = 10, \; \sum_{i=1}^{n} Y_i = 28, \; \sum_{i=1}^{n} X_i^2 = 30, \; \sum_{i=1}^{n} Y_i^2 = 198, \; \sum_{i=1}^{n} X_i Y_i = (4x8) + \ldots + (1x6) = 73.$$

and $$r = \frac{73 - \frac{(10)(28)}{4}}{\left[[30 - \frac{(10)(10)}{4}] \; [198 - \frac{(28)(28)}{4}] \right]^{1/2}} = 0.949$$

Since r is close to 1, there is a strong linear relationship between these two contaminants.

2.3 GRAPHICAL REPRESENTATIONS

2.3.1 Histogram/Frequency Plots

Two of the oldest methods for summarizing data distributions are the frequency plot (Figure 2.3-1) and the histogram (Figure 2.3-2). Both the histogram and the frequency plot use the same basic principles to display the data: dividing the data range into units, counting the number of points within the units, and displaying the data as the height or area within a bar graph. There are slight differences between the histogram and the frequency plot. In the frequency plot, the relative height of the bars represents the relative density of the data. In a histogram, the area within the bar represents the relative density of the data. The difference between the two plots becomes more distinct when unequal box sizes are used.

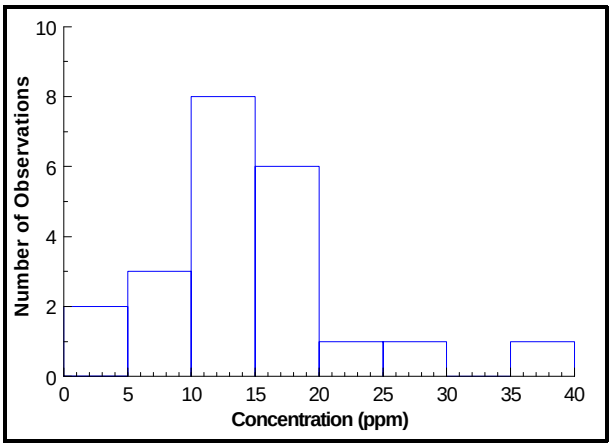


Figure 2.3-1. Example of a Frequency Plot

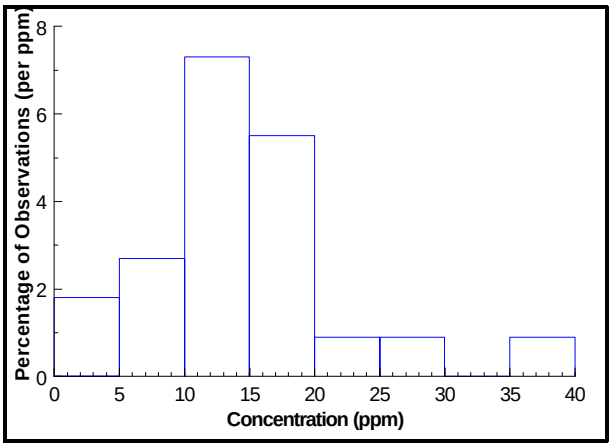


Figure 2.3-2. Example of a Histogram

The histogram and frequency plot provide a means of assessing the symmetry and variability of the data. If the data are symmetric, then the structure of these plots will be symmetric around a central point such as a mean. The histogram and frequency plots will generally indicate if the data are skewed and the direction of the skewness.

Directions for generating a histogram and a frequency plot are contained in Box 2.3-1 and an example is contained in Box 2.3-2. When plotting a histogram for a continuous variable (e.g., concentration), it is necessary to decide on an endpoint convention; that is, what to do with cases that fall on the boundary of a box. With discrete variables, (e.g., family size) the intervals can be centered in between the variables. For the family size data, the intervals can span between 1.5 and 2.5, 2.5 and 3.5, and so on, so that the whole numbers that relate to the family size can be centered within the box. The visual impression conveyed by a histogram or a frequency plot can be quite sensitive to the choice of interval width. The choice of the number of intervals determines whether the histogram shows more detail for small sections of the data or whether the data will be displayed more simply as a smooth overview of the distribution.

Box 2.3-1: Directions for Generating a Histogram and a Frequency Plot

Let X_1, X_2, ..., X_n represent the n data points. To develop a histogram or a frequency plot:

STEP 1: Select intervals that cover the range of observations. If possible, these intervals should have equal widths. A rule of thumb is to have between 7 to 11 intervals. If necessary, specify an endpoint convention, i.e., what to do with cases that fall on interval endpoints.

STEP 2: Compute the number of observations within each interval. For a frequency plot with equal interval sizes, the number of observations represents the height of the boxes on the frequency plot.

STEP 3: Determine the horizontal axis based on the range of the data. The vertical axis for a frequency plot is the number of observations. The vertical axis of the histogram is based on percentages.

STEP 4: For a histogram, compute the percentage of observations within each interval by dividing the number of observations within each interval (Step 3) by the total number of observations.

STEP 5: For a histogram, select a common unit that corresponds to the x-axis. Compute the number of common units in each interval and divide the percentage of observations within each interval (Step 4) by this number. This step is only necessary when the intervals (Step 1) are not of equal widths.

STEP 6: Using boxes, plot the intervals against the results of Step 5 for a histogram or the intervals against the number of observations in an interval (Step 2) for a frequency plot.

Box 2.3-2: Example of Generating a Histogram and a Frequency Plot

Consider the following 22 samples of a contaminant concentration (in ppm): 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.

STEP 1: This data spans 0 - 40 ppm. Equally sized intervals of 5 ppm will be used: 0 - 5 ppm; 5 - 10 ppm; etc. The endpoint convention will be that values are placed in the highest interval containing the value. For example, a value of 5 ppm will be placed in the interval 5 - 10 ppm instead of 0 - 5 ppm.

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

STEP 3: The horizontal axis for the data is from 0 to 40 ppm. The vertical axis for the frequency plot is from 0 - 10 and the vertical axis for the histogram is from 0% - 10%.

STEP 4: There are 22 observations total, so the number observations shown in the table below will be divided by 22. The results are shown in column 3 of the table below.

STEP 5: A common unit for this data is 1 ppm. In each interval there are 5 common units so the percentage of observations (column 3 of the table below) should be divided by 5 (column 4).

STEP 6: The frequency plot is shown in Figure 2.3-1 and the histogram is shown in Figure 2.3-2.

Interval	# of Obs in Interval	% of Obs in Interval	% of Obs per ppm
0 - 5 ppm	2	9.10	1.8
5 - 10 ppm	3	13.60	2.7
10 - 15 ppm	8	36.36	7.3
15 - 20 ppm	6	27.27	5.5
20 - 25 ppm	1	4.55	0.9
25 - 30 ppm	1	4.55	0.9
30 - 35 ppm	0	0.00	0.0
35 - 40 ppm	1	4.55	0.9

2.3.2 Stem-and-Leaf Plot

The stem-and-leaf plot is used to show both the numerical values themselves and information about the distribution of the data. It is a useful method for storing data in a compact form while, at the same time, sorting the data from smallest to largest. A stem-and-leaf plot can be more useful in analyzing data than a histogram because it not only allows a visualization of the data distribution, but enables the data to be reconstructed and lists the observations in the order of magnitude. However, the stem-and-leaf plot is one of the more subjective visualization techniques because it requires the analyst to make some arbitrary choices regarding a partitioning of the data. Therefore, this technique may require some practice or trial and error before a useful plot can be created. As a result, the stem-and-leaf plot should only be used to develop a picture of the data and its characteristics. Directions for constructing a stem-and-leaf plot are given in Box 2.3-3 and an example is contained in Box 2.3-4.

Each observation in the stem-and-leaf plot consist of two parts: the stem of the observation and the leaf. The stem is generally made up of the leading digit of the numerical values while the leaf is made up of trailing digits in the order that corresponds to the order of magnitude from left to right. The stem is displayed on the vertical axis and the data points make up the leaves. Changing the stem can be accomplished by increasing or decreasing the digits that are used, dividing the groupings of one stem (i.e., all numbers which start with the numeral 6 can be divided into smaller groupings), or multiplying the data by a constant factor (i.e., multiply the data by 10 or 100). Nondetects can be placed in a single stem.

A stem-and-leaf plot roughly displays the distribution of the data. For example, the stem-and-leaf plot of normally distributed data is approximately bell shaped. Since the stem-and-leaf roughly displays the distribution of the data, the plot may be used to evaluate whether the data are skewed or symmetric. The top half of the stem-and-leaf plot will be a mirror image of the bottom half of the stem-and-leaf plot for symmetric data. Data that are skewed to the left will have the bulk of data in the top of the plot and less data spread out over the bottom of the plot.

2.3.3 Box and Whisker Plot

A box and whisker plot or box plot (Figure 2.3-3) is a schematic diagram useful for visualizing important statistical quantities of the data. Box plots are useful in situations where it is not necessary or feasible to portray all the details of a distribution. Directions for generating a box and whiskers plot are contained in Box 2.3-5, and an example is contained in Box 2.3-6.

A box and whiskers plot is composed of a central box divided by a line and two lines extending out from the box called whiskers. The length of the central box indicates the spread of the bulk of the data (the central 50%) while the length of the whiskers show how stretched the tails of the distribution are. The width of the box has no particular meaning; the plot can be made quite narrow without affecting its visual impact. The sample median is displayed as a line through the box and the sample mean is displayed using a '+' sign. Any unusually small or large data points are displayed by a '*' on the plot. A box and whiskers plot can be used to assess the symmetry of the data. If the distribution is symmetrical, then the box is divided in two equal halves by the median, the whiskers will be the same length and the number of extreme data points will be distributed equally on either end of the plot.

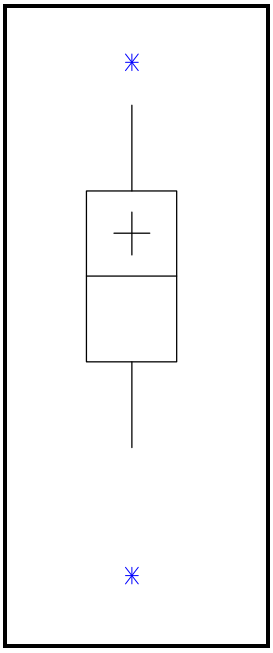


Figure 2.3-3. Example of a Box and Whisker Plot

Box 2.3-3: Directions for Generating a Stem and Leaf Plot

Let $X_1, X_2, ..., X_n$ represent the n data points. To develop a stem-and-leaf plot, complete the following steps:

STEP 1: Arrange the observations in ascending order. The ordered data is usually labeled (from smallest to largest) $X_{(1)}, X_{(2)}, ..., X_{(n)}$.

STEP 2: Choose either one or more of the leading digits to be the stem values. As an example, for the value 16, 1 could be used as the stem as it is the leading digit.

STEP 3: List the stem values from smallest to largest at the left (along a vertical axis). Enter the leaf (the remaining digits) values in order from lowest to highest to the right of the stem. Using the value 16 as an example, if the 1 is the stem then the 6 will be the leaf.

Box 2.3-4: Example of Generating a Stem and Leaf Plot

Consider the following 22 samples of trifluorine (in ppm): 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.

STEP 1: Arrange the observations in ascending order: <4, <4, 5.2, 7.7, 8.9, 10.2, 10.9, 11.6, 12.4, 12.4, 14.7, 14.7, 14.9, 15.2, 16.5, 17.4, 17.7, 19.1, 22.8, 28.6, 35.5.

STEP 2: Choose either one or more of the leading digits to be the stem values. For the above data, using the first digit as the stem does not provide enough detail for analysis. Therefore, the first digit will be used as a stem; however, each stem will have two rows, one for the leaves 0 - 4, the other for the leaves 5 - 9.

STEP 3: List the stem values at the left (along a vertical axis) from smallest to largest. Enter the leaf (the remaining digits) values in order from lowest to highest to the right of the stem. The first digit of the data was used as the stem values; however, each stem value has two leaf rows.

```
0 (0, 1, 2, 3, 4)    | <4   <4
0 (5, 6, 7, 8, 9)    | 5.2  7.7  8.9
1 (0, 1, 2, 3, 4)    | 0.2  0.9  1.6  2.4  2.4  4.7  4.7  4.9
1 (5, 6, 7, 8, 9)    | 5.2  6.5  7.4  7.7  9.1
2 (0, 1, 2, 3, 4)    | 2.8
2 (5, 6, 7, 8, 9)    | 8.6
3 (0, 1, 2, 3, 4)    |
3 (5, 6, 7, 8, 9)    | 5.5
```

Note: If nondetects are present, place them first in the ordered list, using a symbol such as <L. If multiple detection limits were used, place the nondetects in increasing order of detection limits, using symbols such as <L1, <L2, etc. If the first stem extends from zero to a value above the detection limit, then nondetects can be placed in this interval, as shown in the example above. Otherwise, special intervals dedicated to nondetects can be used.

Box 2.3-5: Directions for Generating a Box and Whiskers Plot

STEP 1: Set the vertical scale of the plot based on the maximum and minimum values of the data set. Select a width for the box plot keeping in mind that the width is only a visualization tool. Label the width w; the horizontal scale then ranges from -½W to ½W.

STEP 2: Compute the upper quartile (Q(.75), the 75th percentile) and the lower quartile (Q(.25), the 25th percentile) using Box 2.2-1. Compute the sample mean and median using Box 2.2-2. Then, compute the interquartile range (IQR) where IQR = Q(.75) - Q(.25).

STEP 3: Draw a box through points (-½W, Q(.75)), (-½W, Q (.25)), (½W, Q(.25)) and (½W, Q(.75)). Draw a line from (½W, Q(.5)) to (-½W, Q(.5)) and mark point (0,$\overline{X}$) with (+).

STEP 4: Compute the upper end of the top whisker by finding the largest data value X less than Q(.75) + 1.5(Q(.75) - Q(.25)). Draw a line from (0, Q(.75)) to (0, X).

Compute the lower end of the bottom whisker by finding the smallest data value Y greater than Q(.25) - 1.5(Q(.75) - Q(.25)). Draw a line from (0, Q(.25)) to (0, Y).

STEP 5: For all points X* > X, place an asterisk (*) at the point (0, X*).

For all points Y* < Y, place an asterisk (*) at the point (0, Y*).

Box 2.3-6. Example of a Box and Whiskers Plot

Consider the following 22 samples of trifluorine (in ppm) listed in order from smallest to largest: 4.0, 6.1, 9.8, 10.7, 10.8, 11.5, 11.6, 12.4, 12.4, 14.6, 14.7, 14.7, 16.5, 17, 17.5, 20.6, 20.8, 25.7, 25.9, 26.5, 32.0, and 35.5.

STEP 1: The data ranges from 4.0 to 35.5 ppm. This is the range of the vertical axis. Arbitrarily, a width of 4 will be used for the horizontal axis.

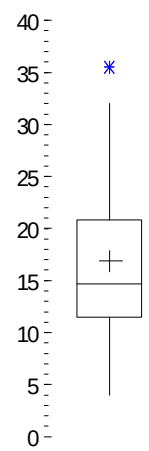
STEP 2: Using the formulas in Box 2.2-2, the sample mean = 16.87 and the median = 14.70. Using Box 2.2-1, Q(.75) = 20.8 and Q(.25) = 11.5. Therefore, IQR = 20.8 - 11.5 = 9.3.

STEP 3: In the figure, a box has been drawn through points (-2, 20.8), (-2, 11.5), (2, 11.5), (2, 20.8). A line has been drawn from (-2 , 14.7) to (2, 14.7), and the point (0, 16.87) has been marked with a '+' sign.

STEP 4: Q(.75) + 1.5(9.3) = 34.75. The closest data value to this number, but less than it, is 32.0. Therefore, a line has been drawn in the figure from (0, 20.8) to (0, 32.0).

Q(.25) - 1.5(9.3) = -2.45. The closest data value to this number, but greater than it, is 4.0. Therefore, a line has been drawn in the figure from (0, 4) to (0, 11.5).

STEP 5: There is only 1 data value greater than 32.0 which is 35.5. Therefore, the point (0, 35.5) has been marked with an asterisk. There are no data values less than 4.0.



2.3.4 Ranked Data Plot

A ranked data plot is a useful graphical representation that is easy to construct, easy to interpret, and makes no assumptions about a model for the data. The analyst does not have to make any arbitrary choices regarding the data to construct a ranked data plot (such as cell sizes for a histogram). In addition, a ranked data plot displays every data point; therefore, it is a graphical representation of the data instead of a summary of the data. Directions for developing a ranked data plot are given in Box 2.3-7 and an example is given in Box 2.3-8.

A ranked data plot is a plot of the data from smallest to largest at evenly spaced intervals (Figure 2.3-4). This graphical representation is very similar to the quantile plot described in section 2.3.5. A ranked data plot is marginally easier to generate than a quantile plot; however, a ranked data plot does not contain as much information as a quantile plot. Both plots can be used to determine the density of the data points and the skewness of the data; however, a quantile plot contains information on the quartiles of the data whereas a ranked data plot does not.

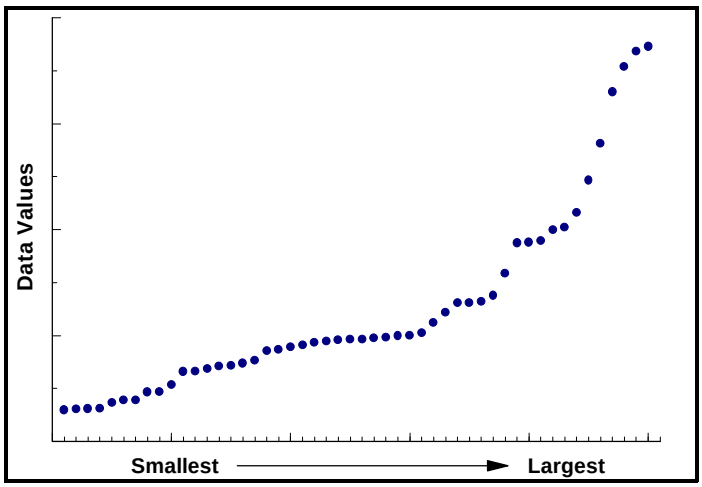


Figure 2.3-4. Example of a Ranked Data Plot

A ranked data plot can be used to determine the density of the data values, i.e., if all the data values are close to the center of the data with relatively few values in the tails or if there is a large amount of values in one tail with the rest evenly distributed. The density of the data is displayed through the slope of the graph. A large amount of data values has a flat slope, i.e., the graph rises slowly. A small amount of data values has a large slope, i.e., the graph rises quickly. Thus the analyst can determine where the data lie, either evenly distributed or in large clusters of points. In Figure 2.3-4, the data rises slowly up to a point where the slope increases and the graph rises relatively quickly. This means that there is a large amount of small data values and relatively few large data values.

A ranked data plot can be used to determine if the data are skewed or if they are symmetric. A ranked data plot of data that are skewed to the right extends more sharply at the top giving the graph a convex shape. A ranked data plot of data that are skewed to the left increases sharply near the bottom giving the graph a concave shape. If the data are symmetric, then the top portion of the graph will stretch to upper right corner in the same way the bottom portion of the graph stretches to lower left, creating a s-shape. Figure 2.3-4 shows a ranked data plot of data that are skewed to the right.

Box 2.3-7: Directions for Generating a Ranked Data Plot

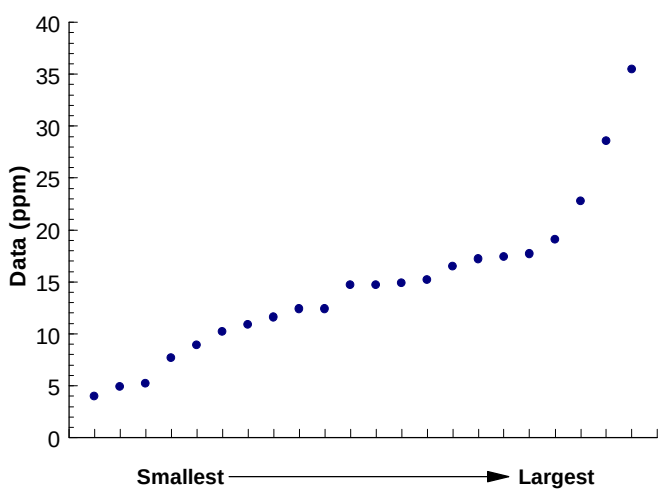
Let $X_1, X_2, ..., X_n$ represent the n data points. Let $X_{(i)}$, for i=1 to n, be the data listed in order from smallest to largest so that $X_{(1)}$ (i = 1) is the smallest, $X_{(2)}$ (i = 2) is the second smallest, and $X_{(n)}$ (i = n) is the largest. To generate a ranked data plot, plot the ordered X values at equally spaced intervals along the horizontal axis.

Box 2.3-8: Example of Generating a Ranked Data Plot

Consider the following 22 samples of triflourine (in ppm): 17.7, 17.4, 22.8, 35.5, 28.6, 17.2 19.1, 4.9, 7.2, 4.0, 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. The data listed in order from smallest to largest $X_{(i)}$ along with the ordered number of the observation (i) are:

i	$X_{(i)}$	i	$X_{(i)}$
1	4.0	12	14.7
2	4.9	13	14.9
3	5.2	14	15.2
4	7.7	15	16.5
5	8.9	16	17.2
6	10.2	17	17.4
7	10.9	18	17.7
8	11.6	19	19.1
9	12.4	20	22.8
10	12.4	21	28.6
11	14.7	22	35.5

A ranked data plot of this data is a plot of the pairs (i, $X_{(i)}$). This plot is shown below:



2.3.5 Quantile Plot

A quantile plot (Figure 2.3-5) is a graphical representation of the data that is easy to construct, easy to interpret, and makes no assumptions about a model for the data. The analyst does not have to make any arbitrary choices regarding the data to construct a quantile plot (such as cell sizes for a histogram). In addition, a quantile plot displays every data point; therefore, it is a graphical representation of the data instead of a summary of the data.

A quantile plot is a graph of the quantiles (section 2.2.1) of the data. The basic quantile plot is visually identical to a ranked data plot except its horizontal axis varies from 0.0 to 1.0, with each point plotted according to the fraction of the points it exceeds. This allows the addition of vertical lines indicating the quartiles or, any other quantiles of interest. Directions for developing a quantile plot are given in Box 2.3-9 and an example is given in Box 2.3-10.

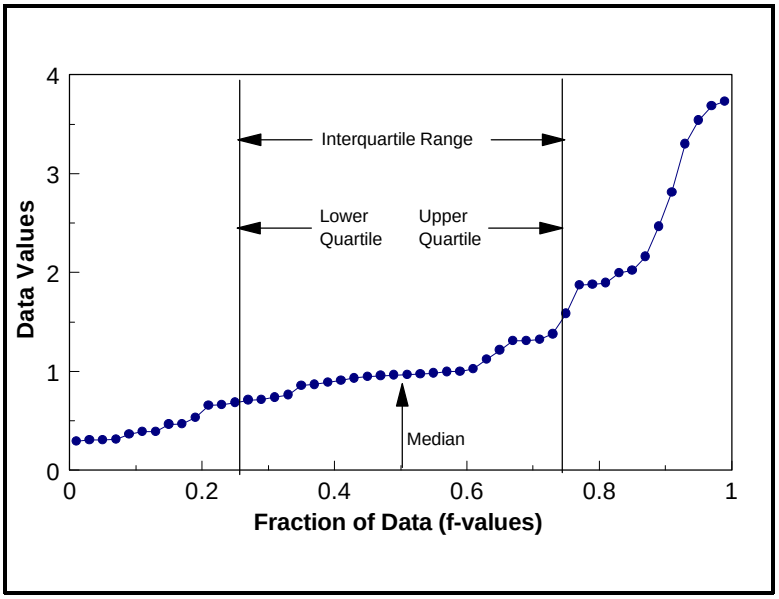


Figure 2.3-5. Example of a Quantile Plot of Skewed Data

A quantile plot can be used to read the quantile information such as the median, quartiles, and the interquartile range. In addition, the plot can be used to determine the density of the data points, e.g., are all the data values close to the center with relatively few values in the tails or are there a large amount of values in one tail with the rest evenly distributed? The density of the data is displayed through the slope of the graph. A large amount of data values has a flat slope, i.e., the graph rises slowly. A small amount of data values has a large slope, i.e., the graph rises quickly. A quantile plot can be used to determine if the data are skewed or if they are symmetric. A quantile plot of data that are skewed to the right is steeper at the top right than the bottom left, as in Figure 2.3-5. A quantile plot of data that are skewed to the left increases sharply near the bottom left of the graph. If the data are symmetric then 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.

Box 2.3-9: Directions for Generating a Quantile Plot

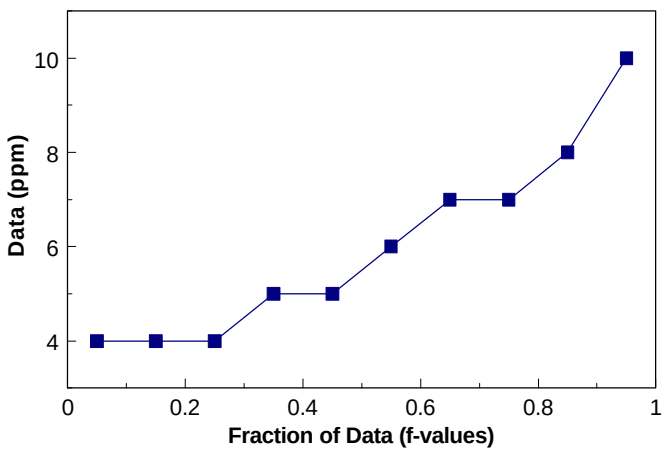
Let $X_1, X_2, ..., X_n$ represent the n data points. To obtain a quantile plot, let $X_{(i)}$, for i = 1 to n, be the data listed in order from smallest to largest so that $X_{(1)}$ (i = 1) is the smallest, $X_{(2)}$ (i = 2) is the second smallest, and $X_{(n)}$ (i = n) is the largest. For each i, compute the fraction $f_i = (i - 0.5)/n$. The quantile plot is a plot of the pairs $(f_i, X_{(i)})$, with straight lines connecting consecutive points.

Box 2.3-10: Example of Generating a Quantile Plot

Consider the following 10 data points: 4 ppm, 5 ppm, 6 ppm, 7 ppm, 4 ppm, 10 ppm, 4 ppm, 5 ppm, 7 ppm, and 8 ppm. The data ordered from smallest to largest, $X_{(i)}$, 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 $(f_i, X_{(i)})$ are then plotted to yield the following quantile plot:



Note that the graph curves upward; therefore, the data appear to be skewed to the right.

2.3.6 Normal Probability Plot (Quantile-Quantile Plot)

There are two types of quantile-quantile plots or q-q plots. The first type, an empirical quantile-quantile plot (section 2.3.7.4), involves plotting the quantiles of two data variables against each other. The second type of a quantile-quantile plot, a theoretical quantile-quantile plot, involves graphing the quantiles of a set of data against the quantiles of a specific distribution. The following discussion will focus on the most common of these plots for environmental data, the normal probability plot (the normal q-q plot); however, the discussion holds for other q-q plots. The normal probability plot is used to roughly determine how well the data set is modeled by a normal distribution. Formal tests are contained in Chapter 4, section 2. Directions for developing a normal probability plot are given in Box 2.3-11 and an example is given in Box 2.3-12.

A normal probability plot is the graph of the quantiles of a data set against the quantiles of the normal distribution using normal probability graph paper (Figure 2.3-6). If the graph is linear, the data may be normally distributed. If the graph is not linear, the departures from linearity give important information about how the data distribution deviates from a normal distribution.

If the graph of the normal probability plot is not linear, the graph may be used to determine the degree of symmetry (or asymmetry) displayed by the data. If the data are skewed to the right, the graph is convex. If the data are skewed to the left, the graph is concave. If the data in the upper tail fall above and the data in the lower tail fall below the quartile line, the data are too slender to be well modeled by a normal distribution, i.e., there are fewer values in the tails of the data set than what is expected from a normal distribution. If the data in the upper tail fall below and the data in the lower tail fall above the quartile line, then the tails of the data are too heavy to be well modeled using a normal distribution, i.e., there are more values in the tails of the data than what is expected from a normal distribution. A normal probability plot can be used to identify 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, ruining the resolution.

Box 2.3-11: Directions for Constructing a Normal Probability Plot

Let $X_1, X_2, ..., X_n$ represent the n data points.

STEP 1: For each data value, compute the absolute frequency, AF_i. The absolute frequency is the number of times each value occurs. For distinct values, the absolute frequency is 1. For non-distinct observations, count the number of times an observation occurs. For example, consider the data 1, 2, 3, 3. The absolute frequency of value 1 is 1 and the absolute frequency of value 2 is 1. The absolute frequency of value 3 is 2 since 3 appears 2 times in the data set.

STEP 2: Compute the cumulative frequencies, CF_i. The cumulative frequency is the number of data points that are less than or equal to X_i, i.e., $CF_i = \sum_{j=1}^{i} AF_j$. Using the data given in step 2, the cumulative frequency for value 1 is 1, the cumulative frequency for value 2 is 2 (1+1), and the cumulative frequency for value 3 is 4 (1+1+2).

STEP 3: Compute $Y_i = 100 \ x \ \frac{CF_i}{(n+1)}$ and plot the pairs (Y_i, X_i) using normal probability paper (Figure 2.3-6). If the graph of these pairs approximately forms a straight line, then the data are probably normally distributed. Otherwise, the data may not be normally distributed.

Box 2.3-12: Example of Normal Probability Plot

Consider the following 15 data points: 5, 5, 6, 6, 8, 8, 9, 10, 10, 10, 10, 10, 12, 14, and 15.

STEP 1: Because the value 5 appears 2 times, its absolute frequency is 2. Similarly, the absolute frequency of 6 is 2, of 8 is 2, of 9 is 1, of 10 is 5, etc. These values are shown in the second column of the table below.

STEP 2: The cumulative frequency of the data value 8 is 6 because there are 2 values of 5, 2 values of 6, and 2 values of 8. The cumulative frequencies are shown in the 3[d] column of the table.

STEP 3: The values $Y_i = 100 \ x \ (\frac{CF_i}{n+1})$ for each data point are shown in column 4 of the table below. A plot of these pairs (Y_i, X_i) using normal probability paper is also shown below.

i	Individual X_i	Absolute Frequency AF_i	Cumulative Frequency CF_i	Y_i
1	5	2	2	12.50
2	6	2	4	25.00
3	8	2	6	37.50
4	9	1	7	43.75
5	10	5	12	75.00
6	12	1	13	81.25
7	14	1	14	87.50
8	15	1	15	93.75

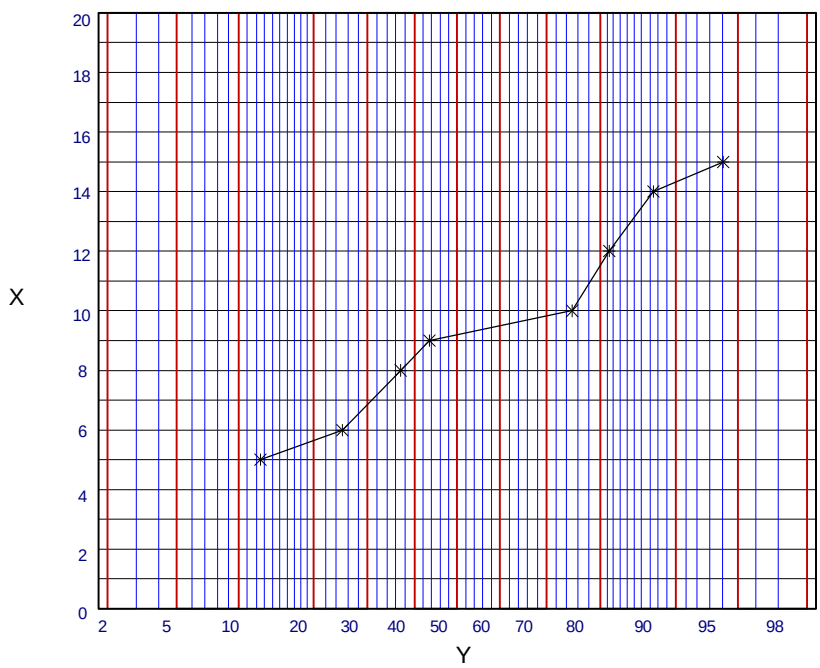


Figure 2.3-6. Normal Probability Paper

2.3.7 Plots for Two or More Variables

Data often consist of measurements of several characteristics (variables) for each sample point in the data set. For example, a data set may consist of measurements of weight, sex, and age for each animal in a sample or may consist of daily temperature readings for several cities. In this case, graphs may be used to compare and contrast different variables. For example, the analyst may wish to compare and contrast the temperature readings for different cities, or different sample points (each containing several variables) such the height, weight, and sex across individuals in a study.

To compare and contrast individual data points, some special plots have been developed to display multiple variables. These plots are discussed in section 2.3.7.1. To compare and contrast several variables, collections of the single variable displays described in previous sections are useful. For example, the analyst may generate box and whisker plots or histograms for each variable using the same axis for all of the variables. Separate plots for each variable may be overlaid on one graph, such as overlaying quantile plots for each variable on one graph. Another useful technique for comparing two variables is to place the stem and leaf plots back to back. In addition, some special plots have been developed to display two or more variables. These plots are described in sections 2.3.7.2 through 2.3.7.4.

2.3.7.1 Plots for Individual Data Points

Since it is difficult to visualize data in more than 2 or 3 dimensions, most of the plots developed to display multiple variables for individual data points involve representing each variable as a distinct piece of a two-dimensional figure. Some such plots include Profiles, Glyphs, and Stars (Figure 2.3-7). These graphical representations start with a specific symbol to represent each data point, then modify the various features of the symbol in proportion to the magnitude of each variable. The proportion of the magnitude is determined by letting the minimum value for each variable be of length 0, the maximum be of length 1. The remaining values of each variable are then proportioned based on the magnitude of each value in relation to the maximum and minimum.

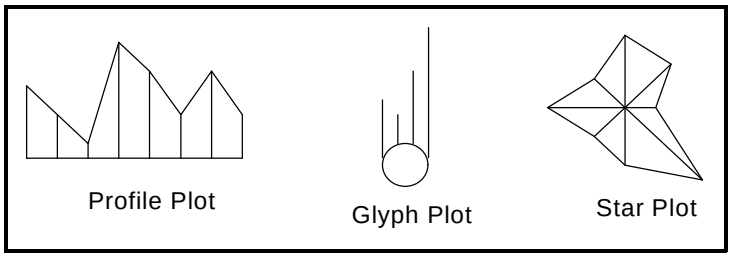


Figure 2.3-7. Example of Graphical Representations of Multiple Variables

A profile plot starts with a line segment of a fixed length. Then lines spaced an equal distance apart and extended perpendicular to the line segment represent each variable. A glyph plot uses a circle of fixed radius. From the perimeter, parallel rays whose sizes are proportional to the magnitude of the variable extend from the top half of the circle. A star plot starts with a point where rays spaced evenly around the circle represent each variable and a polygon is then drawn around the outside edge of the rays.

2.3.7.2 Scatter Plot

For data sets consisting of paired observations where two or more continuous variables are measured for each sampling point, a scatter plot is one of the most powerful tools for analyzing the relationship between two or more variables. Scatter plots are easy to construct for two variables (Figure 2.3-8) and many computer graphics packages can construct 3-dimensional scatter plots. Directions for constructing a scatter plot for two variables are given in Box 2.3-13 along with an example.

A scatter plot clearly shows the relationship between two variables. Both potential outliers from a single variable and potential outliers from the paired variables may be identified on this plot. A scatter plot also displays the correlation between the two variables. Scatter plots of highly linearly correlated variables cluster compactly around a straight line. In addition, nonlinear patterns may be obvious on a scatter plot. For example, consider two variables where one variable is approximately equal to the square of the other. A scatter plot of this data would display a u-shaped (parabolic) curve. Another important feature that can be detected using a scatter plot is any clustering effect among the data.

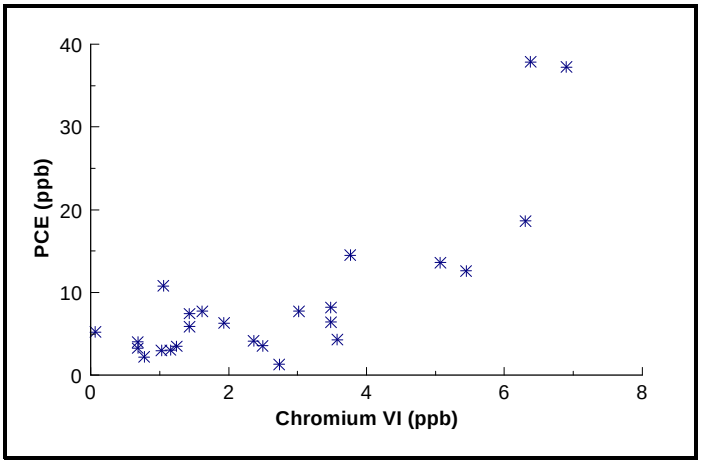


Figure 2.3-8. Example of a Scatter Plot

Box 2.3-13: Directions for Generating a Scatter Plot and an Example

Let $X_1, X_2, ..., X_n$ represent one variable of the n data points and let $Y_1, Y_2, ..., Y_n$ represent a second variable of the n data points. The paired data can be written as (X_i, Y_i) for i = 1, ..., n. To construct a scatter plot, plot the first variable along the horizontal axis and the second variable along the vertical axis. It does not matter which variable is placed on which axis.

Example: A scatter plot will be developed for the data below. PCE values are displayed on the vertical axis and Chromium VI values are displayed on the horizontal axis of Figure 2.3-8.

PCE (ppb)	Chromium VI (ppb)
14.49	3.76
37.21	6.92
10.78	1.05
18.62	6.30
7.44	1.43
37.84	6.38
13.59	5.07
4.31	3.56

PCE (ppb)	Chromium VI (ppb)
2.23	0.77
3.51	1.24
6.42	3.48
2.98	1.02
3.04	1.15
12.60	5.44
3.56	2.49
7.72	3.01

PCE (ppb)	Chromium VI (ppb)
4.14	2.36
3.26	0.68
5.22	0.65
4.02	0.68
6.30	1.93
8.22	3.48
1.32	2.73
7.73	1.61
5.88	1.42

2.3.7.3 Extensions of the Scatter Plot

It is easy to construct a 2-dimensional scatter plot by hand and many software packages can construct a useful 3-dimensional scatter plot. However, with more than 3 variables, it is difficult to construct and interpret a scatter plot. Therefore, several graphical representations have been developed that extend the idea of a scatter plot for data consisting of 2 or more variables.

The simplest of these graphical representations is a coded scatter plot. In this case, all possible pairs of data are given a code and plotted on one scatter plot. For example, consider a data set of 3 variables: variable A, variable B, and variable C. Using the first variable to designate the horizontal axis, the analyst may choose to display the pairs (A, B) using an X, the pairs (A, C) using a Y, and the pairs (B, C) using a Z on one scatter plot. All of the information described above for a scatter plot is also available on a coded scatter plot. However, this method assumes that the ranges of the three variables are comparable and does not provide information on three-way or higher interactions between the variables. An example of a coded scatter plot is given in Figure 2.3-9.

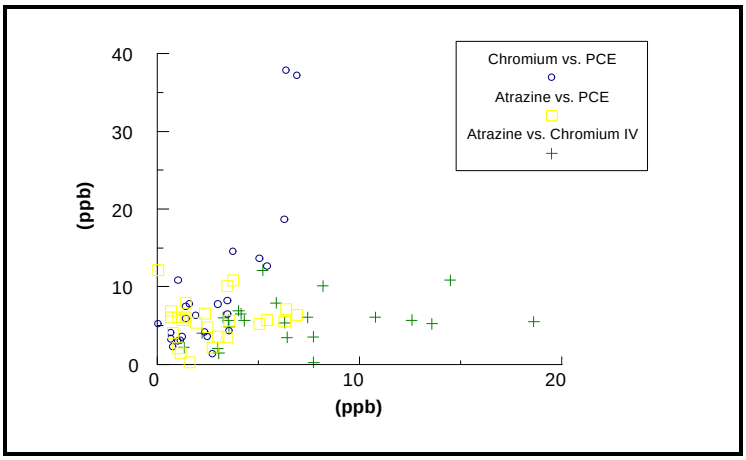


Figure 2.3-9. Example of a Coded Scatter Plot

A parallel coordinate plot also extends the idea of a scatter plot to higher dimensions. The parallel coordinates method employs a scheme where coordinate axes are drawn in parallel (instead of perpendicular). Consider a sample point X consisting of values X_1 for variable 1, X_2 for variable 2, and so on up to X_p for variable p. A parallel coordinate plot is constructed by placing an axis for each of the p variables parallel to each other and plotting X_1 on axis 1, X_2 on axis 2, and so on through X_p on axis p and joining these points with a broken line. This method contains all of the information available on a scatter plot in addition to information on 3-way and higher interactions (e.g., clustering among three variables). However, for p variables one must construct (p+1)/2 parallel coordinate plots in order to display all possible pairs of variables. An example of a parallel coordinate plot is given in Figure 2.3-10.

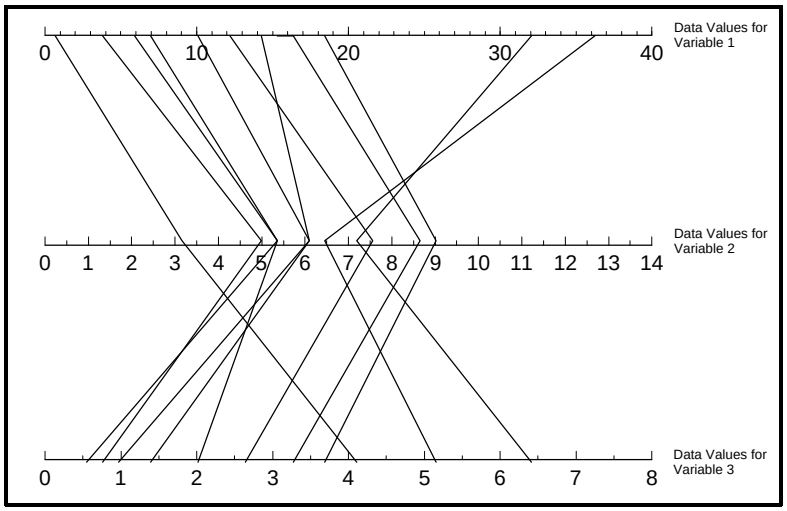


Figure 2.3-10. Example of a Parallel Coordinates Plot

A scatter plot matrix is another useful method of extending scatter plots to higher dimensions. In this case, a scatter plot is developed for all possible pairs of the variables which are then displayed in a matrix format. This method is easy to implement and provides a concise method of displaying the individual scatter plots. However, this method does not contain information on 3-way or higher interactions between variables. An example of a scatter plot matrix is contained in Figure 2.3-11.

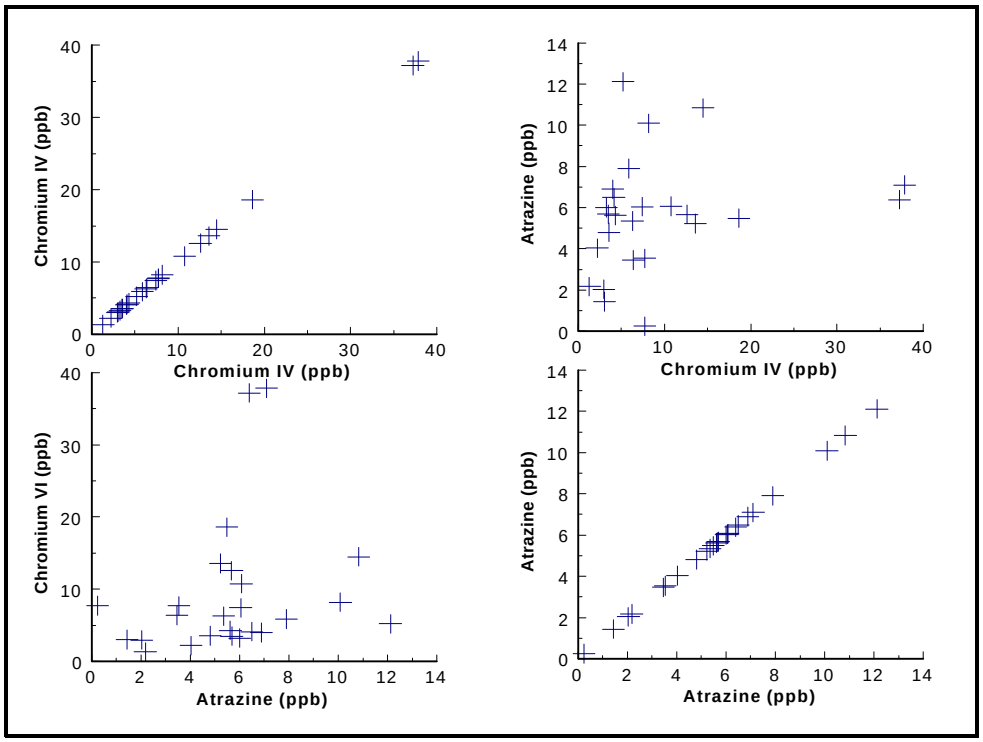


Figure 2.3-11. Example of a Matrix Scatter Plot

2.3.7.4 Empirical Quantile-Quantile Plot

An empirical quantile-quantile (q-q) plot involves plotting the quantiles of two data variables against each other. This plot is used to compare distributions of two or more variables; for example, the analyst may wish to compare the distribution of lead and iron samples from a drinking water well. This plot is similar in concept to the theoretical quantile-quantile plot and yields similar information in regard to the distribution of two variables instead of the distribution of one variable in relation to a fixed distribution. Directions for constructing an empirical q-q plot with an example are given in Box 2.3-14.

An empirical q-q plot is the graph of the quantiles of one variable of a data set against the quantiles of another variable of the data set. This plot is used to determine how well the distribution of the two variables match. If the distributions are roughly the same, the graph is linear or close to linear. If the distributions are not the same, than the graph is not linear. Even if the graph is not linear, the departures from linearity give important information about how the two data distributions differ. For example, a q-q plot can be used to compare the tails of the two data distributions in the same manner a normal probability plot was used to compare the tails of the data to the tails of a normal distribution. In addition, potential outliers (from the paired data) may be identified on this graph.

Box 2.3-14: Directions for Constructing an Empirical Q-Q Plot with an Example

Let $X_1, X_2, ..., X_n$ represent n data points of one variable and let $Y_1, Y_2, ..., Y_m$ represent a second variable of m data points. Let $X_{(i)}$, for i = 1 to n, be the first variable listed in order from smallest to largest so that $X_{(1)}$ (i = 1) is the smallest, $X_{(2)}$ (i = 2) is the second smallest, and $X_{(n)}$ (i = n) is the largest. Let $Y_{(i)}$, for i = 1 to n, be the second variable listed in order from smallest to largest so that $Y_{(1)}$ (i = 1) is the smallest, $Y_{(2)}$ (i = 2) is the second smallest, and $Y_{(m)}$ (i = m) is the largest.

If m = n: If the two variables have the same number of observations, then an empirical q-q plot of the two variables is simply a plot of the ordered values of the variables. Since n=m, replace m by n. A plot of the pairs $(X_{(1)}, Y_{(1)}), (X_{(2)}, Y_{(2)}), ..., (X_{(n)}, Y_{(n)})$ is an empirical quantile-quantile plot.

If n > m: If the two variables have a different number of observations, then the empirical quantile-quantile plot will consist of m (the smaller number) pairs. The empirical q-q plot will then be a plot of the ordered Y values against the interpolated X values. For i = 1, i = 2, ..., i = m, let v = (n/m)(i - 0.5) + 0.5 and separate the result into the integer part and the fractional part, i.e., let v = j + g where j is the integer part and g is the fraction part. If g = 0, plot the pair $(Y_{(i)}, X_{(i)})$. Otherwise, plot the pair $(Y_{(i)}, (1-g)X_{(j)} + gX_{(j+1)})$. A plot of these pairs is an empirical quantile-quantile plot.

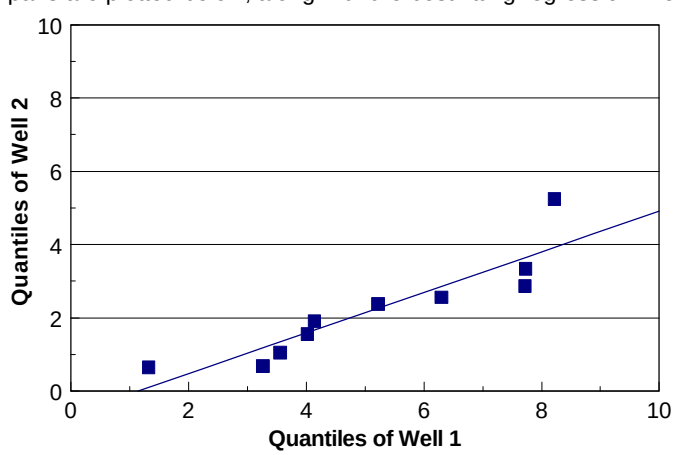
<u>Example</u>: Consider two sets of contaminant readings from two separate drinking water wells at the same site. The data from well 1 are: 1.32, 3.26, 3.56, 4.02, 4.14, 5.22, 6.30, 7.72, 7.73, and 8.22. The data from well 2 are: 0.65, 0.68, 0.68, 1.42, 1.61, 1.93, 2.36, 2.49, 2.73, 3.01, 3.48, and 5.44. An empirical q-q plot will be used to compare the distributions of these two wells. Since there are 10 observations in well 1, and 12 observations in well, the case for n ≠ m will be used. Therefore, for i = 1, 2, ..., 10, compute:

<u>i = 1:</u> $v = \frac{12}{10}(1-.5)+.5 = 1.1$ so j=1 and g=.1. Since g≠0, plot (1.32,(.9).65+(.1).68)=(1.32, 0.653)

<u>i = 2:</u> $v = \frac{12}{10}(2-.5)+.5 = 2.3$ so j=2 and g=.3. Since g≠0, plot (3.26,(.7).68+(.3).68)=(3.26, 0.68)

<u>i = 3:</u> $v = \frac{12}{10}(3-.5)+.5 = 3.5$ so j=3 and g=.5. Since g≠0, plot (3.56,(.5).68+(.5)1.42)=(3.56,1.05)

Continue this process for i =4, 5, 6, 7, 8, 9, and 10 to yield the following 10 data pairs (1.32, 0.653), (3.26, 0.68), (3.56, 1.05), (4.02, 1.553), (4.14, 1.898), (5.22, 2.373), (6.30, 2.562), (7.72, 2.87), (7.73, 3.339), and (8.22, 5.244). These pairs are plotted below, along with the best fitting regression line.



This graph indicates the variables behave roughly the same since there are no substantial deviations from the fitted line.

2.3.8 Plots for Temporal Data

Data collected over specific time intervals (e.g., monthly, biweekly, or hourly) have a temporal component. For example, air monitoring measurements of a pollutant may be collected once a minute or once a day; water quality monitoring measurements of a contaminant level may be collected weekly or monthly. An analyst examining temporal data may be interested in the trends over time, correlation among time periods, and cyclical patterns. Some graphical representations specific to temporal data are the time plot, correlogram, and variogram.

Data collected at regular time intervals are called time series. Time series data may be analyzed using Box-Jenkins modeling and spectral analysis. Both of these methods require a large amount of data collected at regular intervals and are beyond the scope of this guidance. It is recommended that the interested reader consult a statistician.

The graphical representations presented in this section are recommended for all data that have a temporal component regardless of whether formal statistical time series analysis will be used to analyze the data. If the analyst uses a time series methodology, the graphical representations presented below will play an important role in this analysis. If the analyst decides not to use time series methodologies, the graphical representations described below will help identify temporal patterns that need to be accounted for in the analysis of the data.

The analyst examining temporal environmental data may be interested in seasonal trends, directional trends, serial correlation, and stationarity. Seasonal trends are patterns in the data that repeat over time, i.e., the data rise and fall regularly over one or more time periods. Seasonal trends may be large scale, such as a yearly trend where the data show the same pattern of rising and falling over each year, or the trends may be small scale, such as a daily trend where the data show the same pattern for each day. Directional trends are downward or upward trends in the data which is of importance to environmental applications where contaminant levels may be increasing or decreasing. Serial correlation is a measure of the extent to which successive observations are related. If successive observations are related, statistical quantities calculated without accounting for serial correlation may be biased. Finally, another item of interest for temporal data is stationarity (cyclical patterns). Stationary data look the same over all time periods. Directional trends and increasing (or decreasing) variability among the data imply that the data are not stationary.

Temporal data are sometimes used in environmental applications in conjunction with a statistical hypothesis test to determine if contaminant levels have changed. If the hypothesis test does not account for temporal trends or seasonal variations, the data must achieve a “steady state” before the hypothesis test may be performed. Therefore, the data must be essentially the same for comparable periods of time both before and after the hypothesized time of change.

Sometimes multiple observations are taken in each time period. For example, the sampling design may specify selecting 5 samples every Monday for 3 months. If this is the case, the time plot described in section 2.3.8.1 may be used to display the data, display the mean weekly level, display a confidence interval for each mean, or display a confidence interval for each mean with the individual data values. A time plot of all the data can be used to determine if the variability for the different time periods changes. A time plot of the means can be used to determine if the means are possibly changing between time periods. In addition, each time period may be treated as a distinct variable and the methods of section 2.3.7 may be applied.

2.3.8.1 Time Plot

One of the simplest plots to generate that provides a large amount of information is a time plot. A time plot is a plot of the data over time. This plot makes it easy to identify large-scale and small-scale trends over time. Small-scale trends show up on a time plot as fluctuations in smaller time periods. For example, ozone levels over the course of one day typically rise until the afternoon, then decrease, and this process is repeated every day. Larger scale trends, such as seasonal fluctuations, appear as regular rises and drops in the graph. For example, ozone levels tend to be higher in the summer than in the winter so ozone data tend to show both a daily trend and a seasonal trend. A time plot can also show directional trends and increased variability over time. Possible outliers may also be easily identified using a time plot.

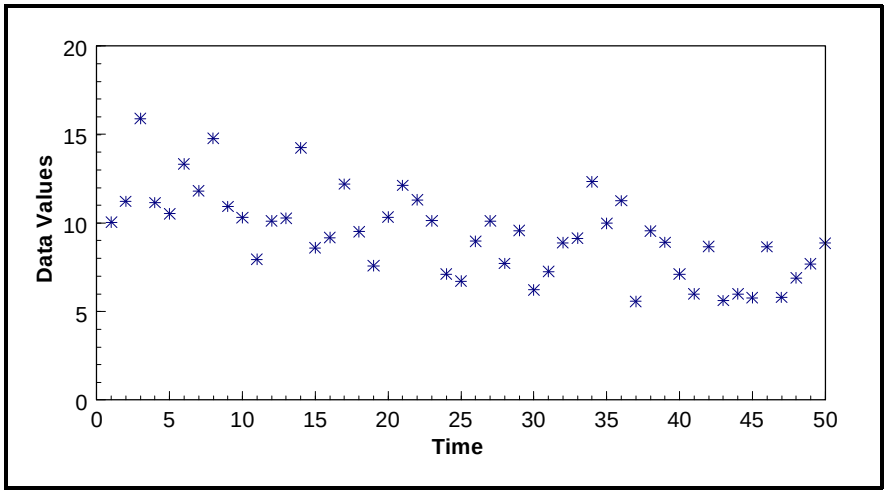


Figure 2.3.12 Example of a Time Plot Showing a Slight Downward Trend

A time plot (Figure 2.3-12) is constructed by numbering the observations in order by time. The time ordering is plotted on the horizontal axis and the corresponding observation is plotted on the vertical axis. The points plotted on a time plot may be joined by lines; however, it is recommended that the plotted points not be connected to avoid creating a false sense of continuity. The scaling of the vertical axis of a time plot is of some importance. A wider scale tends to emphasize large-scale trends, whereas a smaller scale tends to emphasize small-scale trends. Using the ozone example above, a wide scale would emphasize the seasonal component of the data, whereas a smaller scale would tend to emphasize the daily fluctuations. Directions for constructing a time plot are contained in Box 2.3-15 along with an example.

Box 2.3-15: Directions for Generating a Time Plot and an Example

Let $X_1, X_2, ..., X_n$ represent n data points listed in order by time, i.e., the subscript represents the ordered time interval. A plot of the pairs (i, X_i) is a time plot of this data.

Example: Consider the following 50 daily observations (listed in order by day): 10.05, 11.22, 15.9, 11.15, 10.53, 13.33, 11.81, 14.78, 10.93, 10.31, 7.95, 10.11, 10.27, 14.25, 8.6, 9.18, 12.2, 9.52, 7.59, 10.33, 12.13, 11.31, 10.13, 7.11, 6.72, 8.97, 10.11, 7.72, 9.57, 6.23, 7.25, 8.89, 9.14, 12.34, 9.99, 11.26, 5.57, 9.55, 8.91, 7.11, 6.04, 8.67, 5.62, 5.99, 5.78, 8.66, 5.8, 6.9, 7.7, 8.87. By labeling day 1 as 1, day 2 as 2, and so on, a time plot is constructed by plotting the pairs (i, X_i) where i represents the number of the day and X_i represents the concentration level. A time plot of this data is shown in Figure 2.3-12.

2.3.8.2 Plot of the Autocorrelation Function (Correlogram)

Serial correlation is a measure of the extent to which successive observations are related. If successive observations are related, either the data must be transformed or this relationship must be accounted for in the analysis of the data. The correlogram is a plot that is used to display serial correlation when the data are collected at equally spaced time intervals. The autocorrelation function is a summary of the serial correlations of data. The 1st autocorrelation coefficient (r_1) is the correlation between points that are 1 time unit (k_1) apart; the 2nd autocorrelation coefficient (r_2) is the correlation between points that are 2 time units (k_2) apart; etc. A correlogram (Figure 2.3-13) is a plot of the sample autocorrelation coefficients in which the values of k versus the values of r_k are displayed. Directions for constructing a correlogram are contained in Box 2.3-16; example calculations are contained in Box 2.3-17. For large sample sizes, a correlogram is tedious to construct by hand; therefore, software like DataQUEST (QA/G-9D) should be used.

The correlogram is used for modeling time series data and may be used to determine if serial correlation is large enough to create problems in the analysis of temporal data using other methodologies besides formal time series methodologies. A quick method for determining if serial correlation is large is to place horizontal lines at $\pm 2/\sqrt{n}$ on the correlogram (shown as dashed lines on Figure 2.3-13). Autocorrelation coefficients that exceed this value require further investigation.

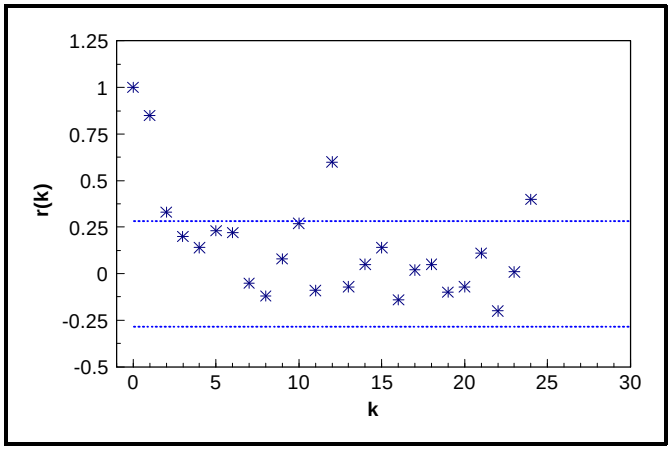


Figure 2.3-13. Example of a Correlogram

In application, the correlogram is only useful for data at equally spaced intervals. To relax this restriction, a variogram may be used instead. The variogram displays the same information as a correlogram except that the data may be based on unequally spaced time intervals. For more information on the construction and uses of the variogram, consult a statistician.

Box 2.3-16: Directions for Constructing a Correlogram

Let $X_1, X_2, ..., X_n$ represent the data points ordered by time for equally spaced time points, i.e., X_1 was collected at time 1, X_2 was collected at time 2, and so on. To construct a correlogram, first compute the sample autocorrelation coefficients. So for k = 0, 1, ..., compute r_k where

$$r_k = \frac{g_k}{g_0} \quad \text{and} \quad g_k = \sum_{t=k+1}^{n} (X_t - \bar{X})(X_{t-k} - \bar{X}).$$

Once the r_k have been computed, a correlogram is the graph (k, r_k) for k = 0, 1, . . . , and so on. As a approximation, compute up to approximately k = n/6. Also, note that $r_0 = 1$. Finally, place horizontal lines at $\pm 2/\sqrt{n}$.

Box 2.3-17: Example Calculations for Generating a Correlogram

A correlogram will be constructed using the following four hourly data points: hour 1: 4.5, hour 2: 3.5, hour 3: 2.5, and hour 4: 1.5. Only four data points are used so that all computations may be shown. Therefore, the idea that no more than n/6 autocorrelation coefficients should be computed will be broken for illustrative purposes. The first step to constructing a correlogram is to compute the sample mean (box 2-2) which is 3 for the 4 points. Then,

$$g_0 = \sum_{t=1}^{4}(y_t-\bar{y})(y_{t-0}-\bar{y}) = \frac{\sum_{t=1}^{4}(y_t-\bar{y})^2}{4} = \frac{(4.5-3)^2+(3.5-3)^2+(2.5-3)^2+(1.5-3)^2}{4} = 1.25$$

$$g_1 = \frac{\sum_{t=2}^{4}(y_t-3)(y_{t-1}-3)}{4} = \frac{(y_2-3)(y_1-3) + (y_3-3)(y_2-3) + (y_4-3)(y_3-3)}{4}$$

$$= \frac{(3.5-3)(4.5-3)+(2.5-3)(3.5-3)+(1.5-3)(2.5-3)}{4} = \frac{1.25}{4} = 0.3125$$

$$g_2 = \frac{\sum_{t=3}^{4}(y_t-3)(y_{t-2}-3)}{4} = \frac{(y_3-3)(y_1-3) + (y_4-3)(y_2-3)}{4}$$

$$= \frac{(2.5-3)(4.5-3) + (1.5-3)(3.5-3)}{4} = \frac{-1.5}{4} = -0.375$$

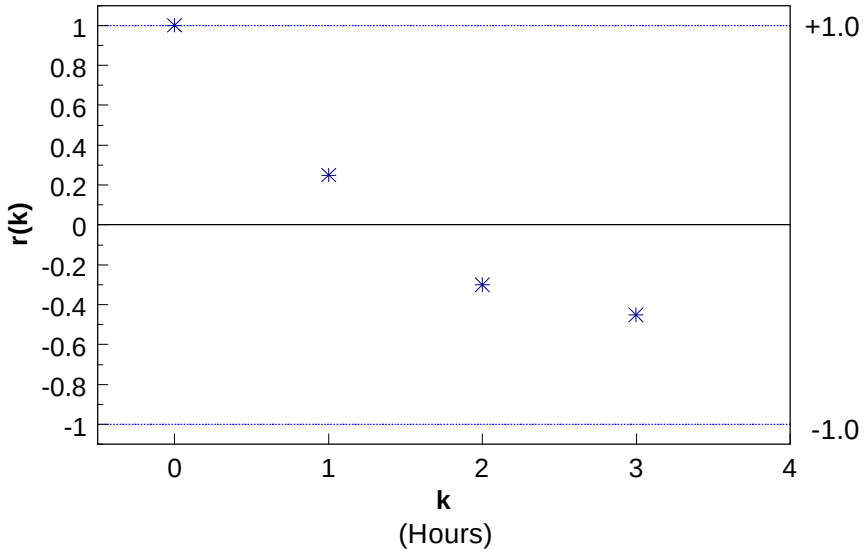
$$g_3 = \frac{\sum_{t=4}^{4}(y_t-3)(y_{t-3}-3)}{4} = \frac{(y_4-3)(y_1-3)}{4} = \frac{(1.5-3)(4.5-3)}{4} = \frac{-2.25}{4} = -0.5625$$

So $r_1 = \frac{g_1}{g_0} = \frac{0.3125}{1.25} = 0.25$, $r_2 = \frac{g_2}{g_0} = \frac{-.375}{1.25} = -0.3$, and $r_3 = \frac{g_3}{g_0} = \frac{-0.5625}{1.25} = -0.45.$

Remember $r_0 = 1$. Thus, the correlogram of these data is a plot of (0, 1) (1, 0.25), (2, -0.3) and (3, -0.45) with two horizontal lines at $\pm 2/\sqrt{4}$ (±1). This graph is shown below.

In this case, it appears that the observations are not serially correlated because all of the correlogram points are within the bounds of $\pm 2/\sqrt{4}$ (±1.0). In Figure 2.3-13, if k represents months, then the correlogram shows a yearly correlation between data points since the points at k=12 and k=24 are out of the bounds of $\pm 2/\sqrt{n}$. This correlation will need to be accounted for when the data are analyzed.

Box 2.3-17: Example Calculations for Generating a Correlogram (Continued)



2.3.8.3 Multiple Observations Per Time Period

Sometimes in environmental data collection, multiple observations are taken for each time period. For example, the data collection design may specify collecting and analyzing 5 samples from a drinking well every Wednesday for three months. If this is the case, the time plot described in section 2.3.8.1. may be used to display the data, display the mean weekly level, display a confidence interval for each mean, or display a confidence interval for each mean with the individual data values. A time plot of all the data will allow the analyst to determine if the variability for the different collection periods varies. A time plot of the means will allow the analyst to determine if the means may possibly be changing between the collection periods. In addition, each collection period may be treated as a distinct variable and the methods described in section 2.3.7 may be applied.

2.3.9 Plots for Spatial Data

The graphical representations of the preceding sections may be useful for exploring spatial data. However, an analyst examining spatial data may be interested in the location of extreme values, overall spatial trends, and the degree of continuity among neighboring locations. Graphical representations for spatial data include postings, symbol plots, correlograms, h-scatter plots, and contour plots.

The graphical representations presented in this section are recommended for all spatial data regardless of whether or not geostatistical methods will be used to analyze the data. The graphical representations described below will help identify spatial patterns that need to be accounted for in the analysis of the data. If the analyst uses geostatistical methods such as kriging to analyze the data, the graphical representations presented below will play an important role in geostatistical analysis.

2.3.9.1 Posting Plots

A posting plot (Figure 2.3-14) is a map of data locations along with corresponding data values. Data posting may reveal obvious errors in data location and identify data values that may be in error. The graph of the sampling locations gives the analyst an idea of how the data were collected (i.e., the sampling design), areas that may have been inaccessible, and areas of special interest to the decision maker which may have been heavily sampled. It is often useful to mark the highest and lowest values of the data to see if there are any obvious trends. If all of the highest concentrations fall in one region of the plot, the analyst may consider some method such as post-stratifying the data (stratification after the data are collected and analyzed) to account for this fact in the analysis. Directions for generating a posting of the data (a posting plot) are contained in Box 2.3-18.

Figure 2.3-14. Example of a Posting Plot

2.3.9.2 Symbol Plots

For large amounts of data, a posting plot may not be feasible and a symbol plot (Figure 2.3-15) may be used. A symbol plot is basically the same as a posting plot of the data, except that instead of posting individual data values, symbols are posted for ranges of the data values. For example, the symbol '0' could

represent all concentration levels less than 100 ppm, the symbol '1' could represent all concentration levels between 100 ppm and 200 ppm, etc. Directions for generating a symbol plot are contained in Box 2.3-18.

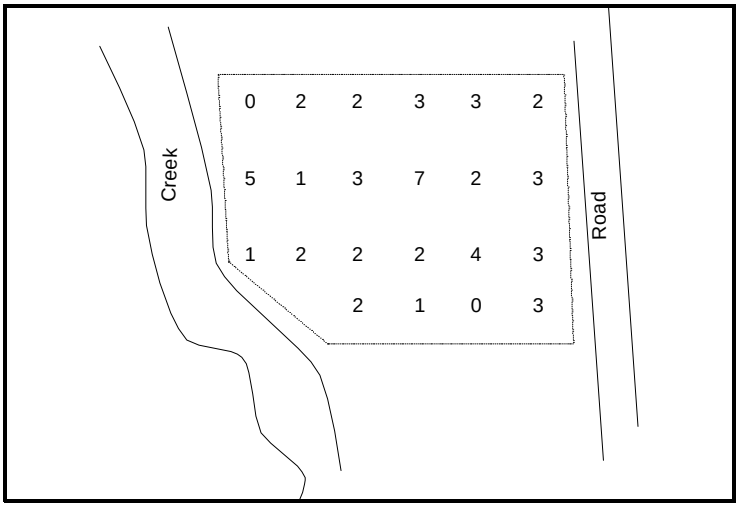


Figure 2.3-15. Example of a Symbol Plot

Box 2.3-18: Directions for Generating a Posting Plot and a Symbol Plot with an Example

On a map of the site, plot the location of each sample. At each location, either indicate the value of the data point (a posting plot) or indicate by an appropriate symbol (a symbol plot) the data range within which the value of the data point falls for that location, using one unique symbol per data range.

Example: The spatial data displayed in the table below contains both a location (Northing and Easting) and a concentration level ([c]). The data range from 4.0 to 35.5 so units of 5 were chosen to group the data:

Range	Symbol	Range	Symbol
0.0 - 4.9	0	20.0 - 24.9	4
5.0 - 9.9	1	25.0 - 29.9	5
10.0 - 14.9	2	30.0 - 34.9	6
15.0 - 19.9	3	35.0 - 39.9	7

The data values with corresponding symbols then become:

Northing	Easting	[c]	Symbol	Northing	Easting	[c]	Symbol
25.0	0.0	4.0	0	15.0	15.0	16.5	3
25.0	5.0	11.6	2	15.0	0.0	8.9	1
25.0	10.0	14.9	2	10.0	5.0	14.7	2
25.0	15.0	17.4	3	10.0	10.0	10.9	2
20.0	0.0	17.7	3	10.0	15.0	12.4	2
20.0	5.0	12.4	2	5.0	0.0	22.8	4
20.0	10.0	28.6	5	5.0	5.0	19.1	3
20.0	15.0	7.7	1	5.0	10.0	10.2	2
15.0	0.0	15.2	3	5.0	15.0	5.2	1
15.0	5.0	35.5	7	0.0	5.0	4.9	0
15.0	10.0	14.7	2	0.0	15.0	17.2	3

The posting plot of this data is displayed in Figure 2.3-14 and the symbol plot is displayed in Figure 2.3-15.

2.3.9.3 Other Spatial Graphical Representations

The two plots described in sections 2.3.9.1 and 2.3.9.2 provide information on the location of extreme values and spatial trends. The graphs below provide another item of interest to the data analyst, continuity of the spatial data. The graphical representations are not described in detail because they are used more for preliminary geostatistical analysis. These graphical representations can be difficult to develop and interpret. For more information on these representations, consult a statistician.

An h-scatterplot is a plot of all possible pairs of data whose locations are separated by a fixed distance in a fixed direction (indexed by h). For example, a h-scatter plot could be based on all the pairs whose locations are 1 meter apart in a southerly direction. A h-scatter plot is similar in appearance to a scatter plot (section 2.3.7.2). The shape of the spread of the data in a h-scatter plot indicates the degree of continuity among data values a certain distance apart in particular direction. If all the plotted values fall close to a fixed line, then the data values at locations separated by a fixed distance in a fixed location are very similar. As data values become less and less similar, the spread of the data around the fixed line increases outward. The data analyst may construct several h-scatter plots with different distances to evaluate the change in continuity in a fixed direction.

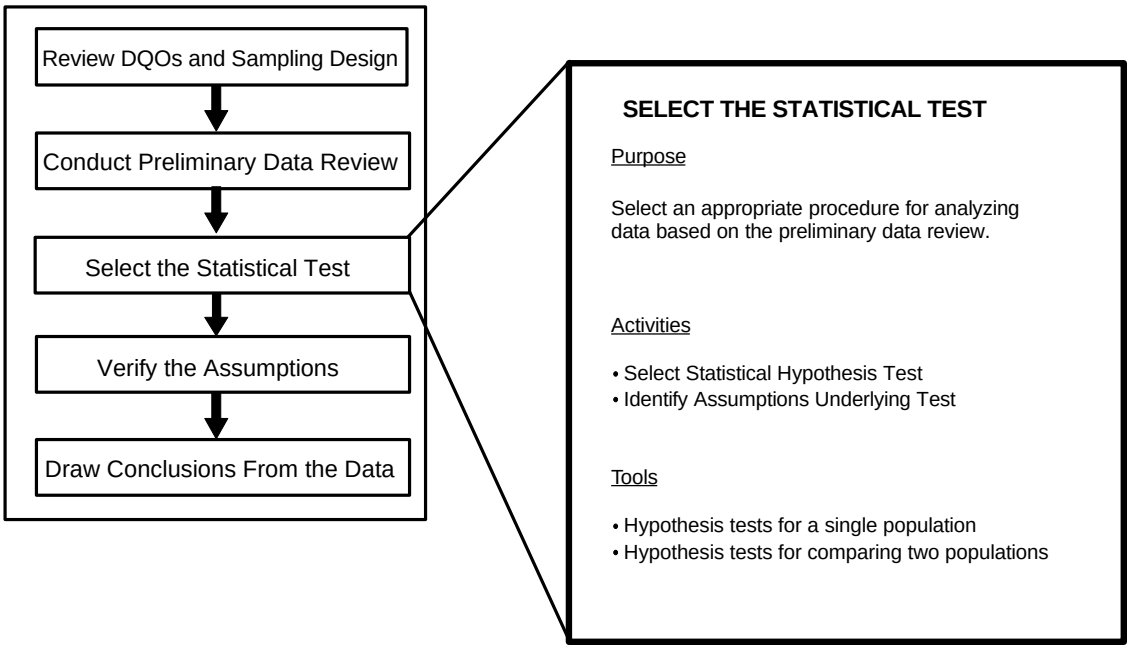
A correlogram is a plot of the correlations of the h-scatter plots. Because the h-scatter plot only displays the correlation between the pairs of data whose locations are separated by a fixed distance in a fixed direction, it is useful to have a graphical representation of how these correlations change for different separation distances in a fixed direction. The correlogram is such a plot which allows the analyst to evaluate the change in continuity in a fixed direction as a function of the distance between two points. A spatial correlogram is similar in appearance to a temporal correlogram (section 2.3.8.2). The correlogram spans opposite directions so that the correlogram with a fixed distance of due north is identical to the correlogram with a fixed distance of due south.

Contour plots are used to reveal overall spatial trends in the data by interpolating data values between sample locations. Most contour procedures depend on the density of the grid covering the sampling area (higher density grids usually provide more information than lower densities). A contour plot gives one of the best overall pictures of the important spatial features. However, contouring often requires that the actual fluctuations in the data values are smoothed so that many spatial features of the data may not be visible. The contour map should be used with other graphical representations of the data and requires expert judgement to adequately interpret the findings.

CHAPTER 3

STEP 3: SELECT THE STATISTICAL TEST

THE DATA QUALITY ASSESSMENT PROCESS



Step 3: Select the Statistical Test

- Select the statistical hypothesis test based on the data user's objectives and the results of the preliminary data review.
  - If the problem involves comparing study results to a fixed threshold, such as a regulatory standard, consider the hypothesis tests in section 3.2.
  - If the problem involves comparing two populations, such as comparing data from two different locations or processes, then consider the hypothesis tests in section 3.3.

- Identify the assumptions underlying the statistical test.
  - List the key underlying assumptions of the statistical hypothesis test, such as distributional form, dispersion, independence, or others as applicable.
  - Note any sensitive assumptions where relatively small deviations could jeopardize the validity of the test results.

STEP 3: SELECT THE STATISTICAL TEST

Parameter	Test	Section	Directions	Example
Mean	One-Sample t-Test	3.2.1.1	Box 3.2-1 Box 3.2-3	Box 3.2-2 Box 3.2-4
	Wilcoxon Signed Rank Test	3.2.1.2	Box 3.2-5 Box 3.2-7	Box 3.2-6
Proportion/ Percentile	One-Sample Proportion Test	3.2.2.1	Box 3.2-8	Box 3.2-9
Two Means	Two-Sample t-Test	3.3.1.1	Box 3.3-1	Box 3.3-2
	Satterthwaite's Two-Sample t-Test	3.3.1.2	Box 3.3-3	Box 3.3-4
Two Proportions/Two Percentiles	Two-Sample Test for Proportions	3.3.2.1	Box 3.3-5	Box 3.3-6
Non-Parametric Comparison of Two Populations	Wilcoxon Rank Sum Test	3.3.3.1	Box 3.3-7 Box 3.3-9	Box 3.3-8
	Quantile Test	3.3.3.2		

Box No. **Page**

3.2-1: Directions for a One-Sample t-Test for Simple and Systematic Random Samples with or without Compositing . . . . . 3.2 - 3
3.2-2: An Example of a One-Sample t-Test for a Simple Random or Composite Sample . . . . . 3.2 - 4
3.2-3: Directions for a One-Sample t-Test for a Stratified Random Sample . . . . . 3.2 - 5
3.2-4: An Example of a One-Sample t-Test for a Stratified Random Sample . . . . . 3.2 - 6
3.2-5: Directions for a Wilcoxon Signed Rank Test for Simple and Systematic Random Samples . . . . 3.2 - 8
3.2-6: An Example of the Wilcoxon Signed Rank Test for a Simple Random Sample . . . . . 3.2 - 9
3.2-7: Directions for the Large Sample Approximation to the Wilcoxon Signed Rank Test for Simple and Systematic Random Samples . . . . . 3.2 - 10
3.2-8: Directions for the One-Sample Test for Proportions for Simple and Systematic Random Samples . . . . . 3.2 - 12
3.2-9: An Example of the One-Sample Test for Proportions for a Simple Random Sample . . . . . 3.2 - 13
3.3-1: Directions for the Student's Two-Sample t-Test (Equal Variances) for Simple and Systematic Random Samples . . . . . 3.3 - 3
3.3-2: An Example of a Student's Two-Sample t-Test (Equal Variances) . . . . . 3.3 - 4
3.3-3: Directions for Satterthwaite's t-Test (Unequal Variances) for Simple and Systematic Random Samples . . . . . 3.3 - 5
3.3-4: An Example of Satterthwaite's t-Test for Simple and Systematic Random Samples . . . . . 3.3 - 6
3.3-5: Directions for a Two-Sample Test for Proportions for Simple and Systematic Samples . . . . . 3.3 - 8
3.3-6: An Example of a Two-Sample Test for Proportions for Simple and Systematic Samples . . . . . 3.3 - 9
3.3-7: Directions for the Wilcoxon Rank Sum Test for Simple and Systematic Samples . . . . . 3.3 - 11
3.3-8: An Example of the Wilcoxon Rank Sum Test for Simple and Systematic Samples . . . . . 3.3 - 12
3.3-9: Directions for the Large Sample Approximation to the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples . . . . . 3.3 - 13

CHAPTER 3
STEP 3: SELECT THE STATISTICAL TEST

3.1 OVERVIEW AND ACTIVITIES

This chapter provides information that the analyst can use in selecting an appropriate statistical hypothesis test that will be used to draw conclusions from the data. A brief review of hypothesis testing is contained in Chapter 1, "Developing DQOs Retrospectively." There are two important outputs from this step: (1) the test itself, and (2) the assumptions underlying the test that determine the validity of conclusions drawn from the test results.

This section describes the two primary activities in this step of the DQA Process. The remaining sections in this chapter contain statistical tests that may be useful for analyzing environmental data. In the one-sample tests discussed in section 3.2, data from a population are compared with an absolute criterion such as a regulatory threshold or action level. In the two-sample tests discussed in section 3.3, data from a population are compared with data from another population (for example, an area expected to be contaminated might be compared with a background area). For each statistical test, this chapter presents its purpose, assumptions, limitations, robustness, and the sequence of steps required to apply the test.

The directions for each hypothesis test given in this chapter are for simple random sampling and randomized systematic sampling designs, except where noted otherwise. If a more complex design is used (such as a stratified design or a composite random sampling design) then different formulas are needed, some of which are contained in this chapter.

3.1.1 Select Statistical Hypothesis Test

If a particular test has been specified either in the DQO Process, the Quality Assurance Project Plan, or by the particular program or study, the analyst should use the results of the preliminary data review to determine if this statistical test is legitimate for the data collected. If the test is not legitimate, the analyst should document why this particular statistical test should not be applied to the data and then select a different test, possibly after consultation with the decision maker. If a particular test has not been specified, the analyst should select a statistical test based on the data user's objectives, preliminary data review, and likely viable assumptions.

3.1.2 Identify Assumptions Underlying the Statistical Test

All statistical tests make assumptions about the data. Parametric tests assume the data have some distributional form (e.g., the t-test assumes normal distribution), whereas nonparametric tests do not make this assumption (e.g., the Wilcoxon test only assumes the data are symmetric but not necessarily normal). However, both parametric and nonparametric tests may assume that the data are statistically independent or that there are no trends in the data. While examining the data, the analyst should always list the underlying assumptions of the statistical hypothesis test, such as distribution, dispersion, or others as applicable.

Another important feature of statistical tests is their sensitivities (nonrobustness) to departures from the assumptions. A statistical procedure is called robust if its performance is not seriously affected by moderate deviations from its underlying assumptions. The analyst should note any sensitive assumptions where relatively small deviations could jeopardize the validity of the test results.

3.2 TESTS OF HYPOTHESES ABOUT A SINGLE POPULATION

A one-sample test involves the comparison of a population parameter (e.g., a mean, percentile, or variance) to a threshold value. Both the threshold value and the population parameter were specified during Step 1: Review DQOs and Sampling Design. In a one-sample test, the threshold value is a fixed number that does not vary. If the threshold value was estimated (and therefore contains variability), a one-sample test is not appropriate. An example of a one-sample test would be to determine if 95% of all companies emitting sulfur dioxide into the air are below a fixed discharge level. For this example, the population parameter is a percentage (proportion) and the threshold value is 95% (.95). Another example is a common Superfund problem that involves comparing the mean contaminant concentration to a risk-based standard. In this case, the risk-based standard (which is fixed) is the threshold value and the statistical parameter is the true mean contaminant concentration level of the site. However, comparing the mean concentration in an area to the mean concentration of a reference area (background) would not be a one-sample test because the mean concentration in the reference area would need to be estimated.

The statistical tests discussed in this section may be used to determine if $\theta \leq \theta_0$ or $\theta > \theta_0$, where θ represents either the population mean, median, a percentile, or a proportion and θ_0 represents the threshold value. Section 3.2.1 discusses tests concerning the population mean, section 3.2.2 discusses tests concerning a proportion or percentile, and section 3.2.2 discusses tests for a median.

3.2.1 Tests for a Mean

A population mean is a measure of the center of the population distribution. It is one of the most commonly used population parameters in statistical hypothesis testing because its distribution is well known for large sample sizes. The hypotheses considered in this section are:

Case 1: H_0: $\mu \leq C$ vs. H_A: $\mu > C$; and

Case 2: H_0: $\mu \geq C$ vs. H_A: $\mu < C$

where C represents a given threshold such as a regulatory level, and μ denotes the (true) mean contaminant level for the population. For example, C may represent the arsenic concentration level of concern. Then if the mean of the population exceeds C, the data user may wish to take action.

The information required for this test (defined in Step 1) includes the null and alternative hypotheses (either Case 1 or Case 2); the gray region, i.e., a value $\mu_1 > C$ for Case 1 or a value $\mu_1 < C$ for Case 2 representing the bound of the gray region; the false positive error rate α at C; the false negative error rate β at μ_1; and any additional limits on decision errors. It may be helpful to label any additional false positive error limits as α_2 at C_2, α_3 at C_3, etc., and to label any additional false negative error limits as β_2 at μ_2, β_3 at μ_3, etc. For example, consider the following decision: determine whether the mean contaminant level at a waste site is greater than 10 ppm. The null hypothesis is H_0: $\mu \geq 10$ ppm and the alternative hypothesis is H_A: $\mu < 10$ ppm. A gray region has been set from 10 to 8 ppm, a false positive error rate of 5% has been set at 10 ppm, and a false negative error rate of 10% has been set at 8 ppm. Thus, C = 10 ppm, $\mu_1 = 8$ ppm, $\alpha = 0.05$, and $\beta = 0.1$. If an additional false negative error rate was set, for example, an error rate of 1% at 4 ppm, then $\beta_2 = .01$ and $\mu_2 = 4$ ppm.

3.2.1.1 The One-Sample t-Test

PURPOSE

Given a random sample of size n (or a composite sample of size n, each composite consisting of k aliquots), the one-sample t-test can be used to test hypotheses involving the mean (μ) of the population from which the sample was selected.

ASSUMPTIONS AND THEIR VERIFICATION

The primary assumptions required for validity of the one-sample t-test are that of a random sample (independence of the data values) and that the sample mean $\bar{x}$ is approximately normally distributed. Because the sample mean and standard deviation are very sensitive to outliers, the t-test should be preceded by a test for outliers (see section 4.4).

Approximate normality of the sample mean follows from approximate normality of the data values. In addition, the Central Limit Theorem states that the sample mean of a random sample from a population with an unknown distribution will be approximately normally distributed provided the sample size is large. This means that although the population distribution from which the data are drawn can be distinctly different from the normal distribution, the distribution of the sample mean can still be approximately normal when the sample size is relatively large. Although preliminary tests for normality of the data can and should be done for small sample sizes, the conclusion that the sample does not follow a normal distribution does not automatically invalidate the t-test, which is robust to moderate violations of the assumption of normality for large sample sizes.

LIMITATIONS AND ROBUSTNESS

The t-test is not robust to outliers because the sample mean and standard deviation are influenced greatly by outliers. The Wilcoxon signed rank test (see section 3.2.1.2) is more robust, but is slightly less powerful. This means that the Wilcoxon signed rank test is slightly less likely to reject the null hypothesis when it is false than the t-test.

The t-test has difficulty dealing with less-than values, e.g., values below the detection limit, compared with tests based on ranks or proportions. Tests based on a proportion above a given threshold (section 3.2.2) are more valid in such a case, if the threshold is above the detection limit. It is also possible to substitute values for below detection-level data (e.g., ½ the detection level) or to adjust the statistical quantities to account for nondetects (e.g., Cohen's Method for normally or lognormally distributed data). See Chapter 4 for more information on dealing with data that are below the detection level.

SEQUENCE OF STEPS

Directions for a one-sample t-test for a simple, systematic, and composite random samples are given in Box 3.2-1 and an example is given in Box 3.2-2. Directions for a one-sample t-test for a stratified random sample are given in Box 3.2-3 and an example is given in Box 3.2-4.

Box 3.2-1: Directions for a One-Sample t-Test for Simple and Systematic Random Samples with or without Compositing

Let $X_1, X_2, \ldots, X_n$ represent the n data points. These could be either n individual samples or n composite samples consisting of k aliquots each. These are the steps for a one-sample t-test for Case 1 (H_0: $\mu \leq C$); modifications for Case 2 (H_0: $\mu \geq C$) are given in braces { }.

STEP 1: Calculate the sample mean $\overline{X}$ (section 2.2.2) and the standard deviation s (section 2.2.3).

STEP 2: Use Table A-1 of Appendix A to find the critical value $t_{1-\alpha}$ such that 100(1-α)% of the t distribution with n - 1 degrees of freedom is below $t_{1-\alpha}$. For example, if α = 0.05 and n = 16, then n-1 = 15 and $t_{1-\alpha}$ = 1.753.

STEP 3: Calculate the sample value $t = (\overline{X} - C) / (s/\sqrt{n})$.

STEP 4: Compare t with $t_{1-\alpha}$.

1) If $t > t_{1-\alpha}$ {$t < -t_{1-\alpha}$}, the null hypothesis may be rejected. Go to Step 6.

2) If $t \not> t_{1-\alpha}$ {$t \not< -t_{1-\alpha}$}, there is not enough evidence to reject the null hypothesis and the false negative error rate should be verified. Go to Step 5.

STEP 5: As the null hypothesis (H_0) was not rejected, calculate either the power of the test or the sample size necessary to achieve the false positive and false negative error rates. To calculate the power, assume that the true values for the mean and standard deviation are those obtained in the sample and use a software package like the Decision Error Feasibility Trial (DEFT) software (EPA G-4D, 1994) or the Data Quality Evaluation Statistical Toolbox (DataQUEST) software (QA/G-9D, 1996) to generate the power curve of the test.

If only one false negative error rate (β) has been specified (at μ_1), it is possible to calculate the sample size which achieves the DQOs, assuming the true mean and standard deviation are equal to the values estimated from the sample, instead of calculating the power of the test. To do this, calculate $m = \dfrac{s^2(z_{1-\alpha} + z_{1-\beta})^2}{(\mu_1 - C)^2} + (0.5)z_{1-\alpha}^2$ where z_p is the pth percentile of the standard normal distribution (Table A-1 of Appendix A). Round m up to the next integer. If $m \leq n$, the false negative error rate has been satisfied. If m > n, the false negative error rate has not been satisfied.

STEP 6: The results of the test may be:

1) the null hypothesis was rejected and it seems that the true mean is less than C {greater than C};

2) the null hypothesis was not rejected and the false negative error rate was satisfied and it seems that the true mean is greater than C {less than C}; or

3) the null hypothesis was not rejected and the false negative error rate was not satisfied and it seems that the true mean is greater than C {less than C} but conclusions are uncertain since the sample size was too small.

Report the results of the test, the sample size, sample mean, standard deviation, t and $t_{1-\alpha}$.

Note: The calculations for the t-test are the same for both simple random or composite random sampling. The use of compositing will usually result in a smaller value of "s" than simple random sampling.

Box 3.2-2: An Example of a One-Sample t-Test for a Simple Random or Composite Sample

Consider the following 9 random (or composite samples each of k aliquots) data points: 82.39 ppm, 103.46 ppm, 104.93 ppm, 105.52 ppm, 98.37 ppm, 113.23 ppm, 86.62 ppm, 91.72 ppm, and 108.21 ppm. This data will be used to test the hypothesis: H_0: $\mu \leq 95$ ppm vs. H_A: $\mu > 95$ ppm. The decision maker has specified a 5% false positive decision error limit (α) at 95 ppm (C), and a 20% false negative decision error limit (β) at 105 ppm (μ_1).

STEP 1: In Boxes 2.3-3 and 2.3-5 of Chapter 2, it was found that

$$\bar{X} = 99.38 \ ppm \quad \text{and} \quad s = 10.41 \ ppm.$$

STEP 2: Using Table A-1 of Appendix A, the critical value of the t distribution with 8 degrees of freedom is $t_{0.95} = 1.86$.

STEP 3:
$$t = \frac{\bar{X} - C}{s/\sqrt{n}} = \frac{99.38 - 95}{10.41/\sqrt{9}} = 1.26$$

STEP 4: Because $1.26 \not> 1.86$, there is not enough evidence to reject the null hypothesis and the false negative error rate should be verified.

STEP 5: Because there is only one false negative error rate, it is possible to use the sample size formula to determine if the error rate has been satisfied. Therefore,

$$m = \frac{s^2(z_{1-\alpha} + z_{1-\beta})^2}{(\mu_1 - C)^2} + (0.5)z_{1-\alpha}^2$$

$$= \frac{10.41^2(1.645 + 0.842)^2}{(95 - 105)^2} + (0.5)(1.645)^2 = 8.049, \text{ i.e., } 9$$

Notice that it is customary to round upwards when computing a sample size. Since m=n, the false negative error rate has been satisfied.

STEP 6: The results of the hypothesis test were that the null hypothesis was not rejected but the false negative error rate was satisfied. Therefore, it seems that the true mean is less than 95 ppm.

Box 3.2-3: Directions for a One-Sample t-Test for a Stratified Random Sample

Let h=1, 2, 3, . . . , L represent the L strata and n_h represent the sample size of stratum h. These steps are for a one-sample t-test for Case 1 (H_0: $\mu \le C$); modifications for Case 2 (H_0: $\mu \ge C$) are given in braces { }.

STEP 1: Calculate the stratum weights (W_h) by calculating the proportion of the volume in stratum h, $W_h = \dfrac{V_h}{\sum_{h=1}^{L} V_h}$ where V_h is the surface area of stratum h multiplied by the depth of sampling in stratum h.

STEP 2: For each stratum, calculate the sample stratum mean $\bar{X}_h = \dfrac{\sum_{i=1}^{n_h} X_{hi}}{n_h}$ and the sample stratum standard error $s_h^2 = \sum_{i=1}^{n_h} \dfrac{(X_{hi} - \bar{X}_h)^2}{n_h - 1}$.

STEP 3: Calculate overall mean $\bar{X}_{ST} = \sum_{h=1}^{L} W_h \bar{X}_h$, and variance $s_{ST}^2 = \sum_{h=1}^{L} W_h^2 \dfrac{s_h^2}{n_h}$.

STEP 4: Calculate the degrees of freedom (dof): $dof = \dfrac{(s_{ST}^2)^2}{\sum_{h=1}^{L} \dfrac{W_h^4 s_h^4}{n_h^2 (n_h - 1)}}$.

Use Table A-1 of Appendix A to find the critical value $t_{1-\alpha}$ so that 100(1-α)% of the t distribution with the above degrees of freedom (rounded to the next highest integer) is below $t_{1-\alpha}$.

STEP 5: Calculate the sample value: $t = \dfrac{\bar{X}_{ST} - C}{\sqrt{s_{ST}}}$

STEP 6: Compare t to $t_{1-\alpha}$. If $t > t_{1-\alpha}$ {$t < -t_{1-\alpha}$}, the null hypothesis may be rejected. Go to Step 8. If $t \not> t_{1-\alpha}$ {$t \not< -t_{1-\alpha}$}, there is not enough evidence to reject the null hypothesis and the false negative error rate should be verified. Go to Step 7.

STEP 7: If the null hypothesis was not rejected, calculate either the power of the test or the sample size necessary to achieve the false positive and false negative error rates (see Step 5, Box 3.2-1).

STEP 8: The results of the test may be:

1) the null hypothesis was rejected so it seems that the true mean is less than C {greater than C};

2) the null hypothesis was not rejected and the false negative error rate was satisfied and it seems that the true mean is greater than C {less than C}; or

3) the null hypothesis was not rejected and the false negative error rate was not satisfied and it seems that the true mean is greater than C {less than C} but conclusions are uncertain since the sample size was too small.

Report the results of the test, as well as the sample size, sample mean, and sample standard deviation for each stratum, the estimated t, the dof, and $t_{1-\alpha}$.

Box 3.2-4: An Example of a One-Sample t-Test for a Stratified Random Sample

Consider a stratified sample consisting of two strata where stratum 1 comprises 10% of the total site surface area and stratum 2 comprises the other 90%, and 40 samples were collected from stratum 1, and 60 samples were collected from stratum 2. For stratum 1, the sample mean is 23 ppm and the sample standard deviation is 18.2 ppm. For stratum 2, the sample mean is 35 ppm, and the sample standard deviation is 20.5 ppm. This information will be used to test the null hypothesis that the overall site mean is greater than or equal to 40 ppm, i.e., H_0: $\mu \geq 40$ ppm (Case 2). The decision maker has specified a 1% false positive decision limit (α) at 40 ppm and a 20% false negative decision error limit (β) at 35 ppm (μ_1).

STEP 1: $W_1 = 10/100 = 0.10$, $W_2 = 90/100 = 0.9$.

STEP 2: From above, $\overline{X}_1 = 23$ ppm, $\overline{X}_2 = 35$ ppm, $s_1 = 18.2$, and $s_2 = 20.5$. This information was developed using the equations in step 2 of Box 3.2-3.

STEP 3: The estimated overall mean concentration is:

$$\overline{X}_{ST} = \sum_{h=1}^{L} W_h \overline{X}_h = W_1\overline{X}_1 + W_2\overline{X}_2 = (.1)(23) + (.9)(35) = 33.8 \ ppm.$$

and the estimated overall variance is:

$$s_{ST}^{\ 2} = \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{n_h} = \frac{(.1)^2(18.2)^2}{40} + \frac{(.9)^2(20.5)^2}{60} = 5.76$$

STEP 4: The approximate degrees of freedom (dof) is:

$$dof = \frac{(s_{ST}^2)^2}{\sum_{h=1}^{L} \frac{W_h^4 s_h^4}{n_h^2(n_h - 1)}} = \frac{(5.76)^2}{\frac{(.1)^4(18.2)^4}{(40)^2 39} + \frac{(.9)^4(20.5)^4}{(60)^2 59}} = 60.8, \text{ i.e., } 61$$

Note how the degrees of freedom has been rounded up to a whole number. Using Table A-1 of Appendix A, the critical value $t_{1-\alpha}$ of the t distribution with 61 dof is approximately 2.39.

STEP 5: Calculate the sample value $t = \dfrac{\overline{X}_{ST} - C}{\sqrt{s_{ST}}} = \dfrac{33.8 - 40}{\sqrt{5.76}} = -2.58$

STEP 6: Because -2.58 < -2.39 the null hypothesis may be rejected.

STEP 7: Because the null hypothesis was rejected, it is concluded that the mean is probably less than 40 ppm. In this example there is no need to calculate the false negative rate as the null hypothesis was rejected and so the chance of making a false negative error is zero by definition.

3.2.1.2 The Wilcoxon Signed Rank (One-Sample) Test for the Mean

PURPOSE

Given a random sample of size n (or composite sample size n, each composite consisting of k aliquots), the Wilcoxon signed rank test can be used to test hypotheses regarding the population mean or median of the population from which the sample was selected.

ASSUMPTIONS AND THEIR VERIFICATION

The Wilcoxon signed rank test assumes that the data constitute a random sample from a symmetric continuous population. (Symmetric means that the underlying population frequency curve is symmetric about its mean/median.) Symmetry is a less stringent assumption than normality since all normal distributions are symmetric, but some symmetric distributions are not normal. The mean and median are equal for a symmetric distribution, so the null hypothesis can be stated in terms of either parameter. Tests for symmetry can be devised which are based on the chi-squared distribution, or a test for normality may be used. If the data are not symmetric, it may be possible to transform the data so that this assumption is satisfied. See Chapter 4 for more information on transformations and tests for symmetry.

LIMITATIONS AND ROBUSTNESS

Although symmetry is a weaker assumption than normality, it is nonetheless a strong assumption. If the data are not approximately symmetric, this test should not be used. For large sample sizes ($n > 50$), the t-test is more robust to violations of its assumptions than the Wilcoxon signed rank test. For small sample sizes, if the data are not approximately symmetric and are not normally distributed, this guidance recommends consulting a statistician before selecting a statistical test or changing the population parameter to the median and applying a different statistical test (section 3.2.3).

The Wilcoxon signed rank test may produce misleading results if many data values are the same. When values are the same, their relative ranks are the same, and this has the effect of diluting the statistical power of the Wilcoxon test. Box 3.2-5 demonstrates the correct method used to break tied ranks. If possible, results should be recorded with sufficient accuracy so that a large number of equal values do not occur. Estimated concentrations should be reported for data below the detection limit, even if these estimates are negative, as their relative magnitude to the rest of the data is of importance.

SEQUENCE OF STEPS

Directions for the Wilcoxon signed rank test for a simple random sample and a systematic simple random sample are given in Box 3.2-5 and an example is given in Box 3.2-6 for samples sizes smaller than 20. For sample sizes greater than 20, the large sample approximation to the Wilcoxon Signed Rank Test should be used. Directions for this test are given in Box 3.2-7.

Box 3.2-5: Directions for the Wilcoxon Signed Rank Test for Simple and Systematic Random Samples

Let $X_1, X_2, \ldots, X_n$ represent the n data points. The following describes the steps for applying the Wilcoxon signed rank test for both Case 1 (H_0: $\mu \leq C$) and Case 2 (H_0: $\mu \geq C$) for a sample size (n) less than 20. If the sample size is greater than or equal to 20, use Box 3.2-7.

STEP 1: If possible, assign values to any measurements below the detection limit. If this is not possible, assign the value "Detection Limit divided by 2" to each value. Then subtract C from each of the n observations X_i to obtain the deviations $d_i = X_i - C$. If any of the deviations are zero delete them and correspondingly reduce the sample size n.

STEP 2: Assign ranks from 1 to n based on ordering the absolute deviations $|d|$ (i.e., magnitude of differences ignoring the sign) from smallest to largest. The rank 1 is assigned to the smallest value, the rank 2 to the second smallest value, and so forth. If there are ties, assign the average of the ranks which would otherwise have been assigned to the tied observations.

STEP 3: Calculate the signed rank for each observation. This signed rank is equal to the rank if the deviation d_i is positive, or equal to the negative rank if the deviation d_i is negative.

STEP 4: For Case 1, calculate the sum R of the ranks with a positive sign.

For Case 2, calculate the sum R of the ranks with a negative sign and take the absolute value of this sum (i.e., ignore the negative sign).

STEP 5: Use Table A-6 of Appendix A to find the critical value w_α.

If $R \geq w_\alpha$, the null hypothesis may be rejected. Go to Step 7.

If $R < w_\alpha$, there is not enough evidence to reject the null hypothesis, and the false negative error rate will need to be verified. Go to Step 6.

STEP 6: If the null hypothesis (H_0) was not rejected, calculate either the power of the test or the sample size necessary to achieve the false positive and false negative error rates using a software package like the DEFT software (EPA G-4D, 1994) or the DataQUEST software (EPA G-4D, 1996). Calculate,

$$m = \frac{s^2(z_{1-\alpha}+z_{1-\beta})^2}{(\mu_1 - C)^2} + (0.5)z_{1-\alpha}^2$$

where z_p is the pth percentile of the standard normal distribution (Table A-1 of Appendix A). Then multiply m by 1.16 to account for loss in efficiency and if this is number is greater than or equal to n, the false negative error rate has been satisfied.

STEP 7: The results of the test may be:

1) the null hypothesis was rejected, and for Case 1, it seems the true mean is greater than C or for Case 2, it seems the true mean is less than C;

2) the null hypothesis was not rejected, the false negative error rate was satisfied, and for Case 1, it seems the true mean is less than C or for Case 2, it seems the true mean is greater than C; or

3) the null hypothesis was not rejected, the false negative error rate was not satisfied, and for Case 1, it seems the true mean is less than C or for Case 2, it seems the true mean is greater than C but the conclusions are uncertain because the sample size was too small.

Box 3.2-6: An Example of the Wilcoxon Signed Rank Test for a Simple Random Sample

Consider the following 10 data points: 974 ppb, 1044 ppb, 1093 ppb, 897 ppb, 879 ppb, 1161 ppb, 839 ppb, 824 ppb, 796 ppb, and one observation below the detection limit of 750 ppb. This data will be used to test the hypothesis: H_0: $\mu \geq 1000$ ppb vs. H_A: $\mu < 1000$ ppb (Case 2). The decision maker has specified a 10% false positive decision error limit (α) at 1000 ppb (C), and a 20% false negative decision error limit (β) at 900 ppb (μ_1).

STEP 1: Assign the value 375 ppb (750 divided by 2) to the data point below the detection limit. Subtract C (1000) from each of the n observations X_i to obtain the deviations $d_i = X_i - 1000$. This is shown in row 2 of the table below.

X_i	974	1044	1093	897	879	1161	839	824	796	375
d_i	-26	+44	+93	-103	-121	+161	-161	-176	-204	-625
$\|d_i\|$	26	44	93	103	121	161	161	176	204	625
rank	1	2	3	4	5	6.5	6.5	8	9	10
s-rank	-1	2	3	-4	-5	6.5	-6.5	-8	-9	-10

STEP 2: Assign ranks from 1 to n based on ordering the absolute deviations $|d|$ (magnitude ignoring any negative sign) from smallest to largest. The absolute deviations are listed in row 3 of the table above. Note that the data have been sorted (rearranged) for clarity so that the absolute deviations are ordered from smallest to largest.

The rank 1 is assigned to the smallest value, the rank 2 to the second smallest value, and so forth. Observations 6 and 7 are ties, therefore, the average (6+7)/2 = 6.5 will be assigned to the two observations. The ranks are shown in row 4.

STEP 3: Calculate the signed rank for each observation. This signed rank is equal to the rank if the deviation d_i is positive, or equal to the negative rank if the deviation d_i is negative. The signed rank is shown in row 5.

STEP 4: Because of the form of the null hypothesis (H_0: $\mu \geq 1000$ ppb), R is the sum of the ranks with <u>negative</u> signs. Since, -1 + -4 + -5 + -6.5 + -8 + -9 + -10 = -43.5, R = 43.5.

STEP 5: Because there are only 10 data points, Table A-6 of Appendix A is used to find the critical value w_α where $\alpha = 0.10$. For this example, $w_{0.10} = 15$. Therefore, since 43.5 > 15, the null hypothesis may be rejected.

STEP 6: The null hypothesis was rejected with a 10% significance level using the Wilcoxon signed rank test (w=15). Therefore, it would seem that the true mean is below 1000 ppb.

Box 3.2-7: Directions for the Large Sample Approximation to the Wilcoxon Signed Rank Test for Simple and Systematic Random Samples

Let $X_1, X_2, \ldots, X_n$ represent the n data points where n is greater than or equal to 20. The following describes the steps for applying the large sample approximation for the Wilcoxon signed rank test for both Case 1 (H_0: $\mu \leq C$) and Case 2 (H_0: $\mu \geq C$).

STEP 1: If possible, assign values to any measurements below the detection limit. If this is not possible, assign the value "Detection Limit divided by 2" to each value. Then subtract C from each of the n observations X_i to obtain the deviations $d_i = X_i - C$. If any of the deviations are zero delete them and correspondingly reduce the sample size n.

STEP 2: Assign ranks from 1 to n based on ordering the absolute deviations $|d|$ (i.e., magnitude of differences ignoring the sign) from smallest to largest. The rank 1 is assigned to the smallest value, the rank 2 to the second smallest value, and so forth. If there are ties, assign the average of the ranks which would otherwise have been assigned to the tied observations.

STEP 3: Calculate the signed rank for each observation. This signed rank is equal to the rank if the deviation d_i is positive, or equal to the negative rank if the deviation d_i is negative.

STEP 4: For Case 1, calculate the sum R of the ranks with a positive sign. For Case 2, calculate the sum R of the ranks with a negative sign and take the absolute value of this sum (i.e., ignore the negative sign). Then calculate:

$$z = \frac{R - \frac{n(n+1)}{4}}{\sqrt{n(n+1)(2n+1)/24}}$$

STEP 5: Use Table A-1 of Appendix A to find the critical value $z_{1-\alpha}$ such that $100(1-\alpha)\%$ of the normal distribution is below $z_{1-\alpha}$. For example, if $\alpha = 0.05$, then $z_{1-\alpha} = 1.645$. If $z > z_{1-\alpha}$, the null hypothesis may be rejected. If $z \not> z_{1-\alpha}$, there is not enough evidence to reject the null hypothesis. Therefore, the false negative error rate will need to be verified.

STEP 6: If the null hypothesis (H_0) was not rejected, calculate either the power of the test or the sample size necessary to achieve the false positive and false negative error rates using a software package like the DEFT software (EPA G-4D, 1994) or the DataQUEST software (EPA G-4D, 1996). Calculate,

$$m = \frac{s^2(z_{1-\alpha}+z_{1-\beta})^2}{(\mu_1 - C)^2} + (0.5)z_{1-\alpha}^2$$

where z_p is the pth percentile of the standard normal distribution (Table A-1 of Appendix A). Then multiply m by 1.16 to account for loss in efficiency and if this value is greater than or equal to n, the false negative error rate has been satisfied.

STEP 7: The results of the test may be:

1) the null hypothesis was rejected, and for Case 1, it seems the true mean is greater than C or for Case 2, it seems the true mean is less than C;

2) the null hypothesis was not rejected, the false negative error rate was satisfied, and for Case 1, it seems the true mean is less than C or for Case 2, it seems the true mean is greater than C; or

3) the null hypothesis was not rejected, the false negative error rate was not satisfied, and for Case 1, it seems the true mean is less than C or for Case 2, it seems the true mean is greater than C but the conclusions are uncertain because the sample size was too small.

3.2.2 Tests for a Proportion or Percentile

This section considers hypotheses concerning population proportions and percentiles. A population proportion is the ratio of the number of elements of a population that has some specific characteristic to the total number of elements. A population percentile represents the percentage of elements of a population having values less than some threshold C. Thus, if x is the 95 th percentile of a population, 95% of the elements of the population have values less than C and 5% of the population have values greater than C.

This section of the guidance covers the following hypothesis: Case 1: H_0: $P \le P_0$ vs. H_A: $P > P_0$ and Case 2: H_0: $P \ge P_0$ vs. H_A: $P < P_0$ where P is a proportion of the population, and P_0 represents a given proportion ($0 \le P_0 \le 1$). Equivalent hypotheses written in terms of percentiles are H_0: the 100Pth percentile is C or larger for Case 1, and H_0: the 100Pth percentile is C or smaller for Case 2. For example, consider the decision to determine whether the 95 th percentile of a container of waste is less than 1 mg/L cadmium. The null hypothesis in this case is H_o: the 95th percentile of cadmium is less than 1 mg/L. Now, instead of considering the population to consist of differing levels of cadmium, consider the population to consist of a binary variable that is '1' if the cadmium level is above 1 mg/L or is '0' if the level is below 1 mg/L. In this case, the hypothesis may be changed to a test for a proportion so that the null hypothesis becomes H_0: $P < .95$ where P represents the proportion of 1's (cadmium levels above 1 mg/L) in the container of waste. Thus, any hypothesis about the proportion of the site below a threshold can be converted to an equivalent hypothesis about percentiles. Therefore, only hypotheses about the proportion of the site below a threshold will be discussed in this section. The information required for this test includes the null and alternative hypotheses, the gray region, the false positive error rate α at P_0, the false negative error rate β at P_1, and any additional limits on decision errors. It may be helpful to label any additional false positive error limits as α_2 at $P_{\alpha 2}$, α_3 at $P_{\alpha 3}$, etc., and any additional false negative error limits as β_2 at $P_{\beta 2}$, β_3 at $P_{\beta 3}$, etc.

3.2.2.1 The One-Sample Proportion Test

PURPOSE

Given a random sample of size n, the one-sample proportion test may be used to test hypotheses regarding a population proportion or population percentile for a distribution from which the data were drawn. Note that for P=.5, this test is also called the Sign test.

ASSUMPTIONS AND THEIR VERIFICATION

The only assumption required for the one-sample proportion test is the assumption of a random sample. To verify this assumption, review the procedures and documentation used to select the sampling points and ascertain that proper randomization has been used.

LIMITATIONS AND ROBUSTNESS

Since the only assumption is that of a random sample, the procedures are valid for any underlying distributional shape. The procedures are also robust to outliers, as long as they do not represent data errors.

SEQUENCE OF STEPS

Directions for the one-sample test for proportions for a simple random sample and a systematic simple random sample are given in Box 3.2-8, an example is given in Box 3.2-9.

Box 3.2-8: Directions for the One-Sample Test for Proportions for Simple and Systematic Random Samples

This box describes the steps for applying the one-sample test for proportions for Case 1 (H_0: $P \le P_0$); modifications for Case 2 (H_0: $P \ge P_0$) are given in braces { }.

STEP 1: Given a random sample $X_1, X_2, \ldots, X_n$ of measurements from the population, let p (small p) denote the proportion of X's that do not exceed C, i.e., p is the number (k) of sample points that are less than or equal to C, divided by the sample size n.

STEP 2: Compute np, and n(1-p). If both np and n(1-p) are greater than or equal to 5, use Steps 3 and 4. Otherwise, consult a statistician as analysis may be complex.

STEP 3: Calculate $z = \dfrac{p - .5/n - P_0}{\sqrt{P_0(1-P_0)/n}}$ for Case 1 or $z = \dfrac{p + .5/n - P_0}{\sqrt{P_0(1-P_0)/n}}$ for Case 2.

STEP 4: Use Table A-1 of Appendix A to find the critical value $z_{1-\alpha}$ such that 100(1-α)% of the normal distribution is below $z_{1-\alpha}$. For example, if $\alpha = 0.05$ then $z_{1-\alpha} = 1.645$.

If $z > z_{1-\alpha}$ {$z < -z_{1-\alpha}$}, the null hypothesis may be rejected. Go to Step 6.

If $z \not> z_{1-\alpha}$ {$z \not< -z_{1-\alpha}$}, there is not enough evidence to reject the null hypothesis. Therefore, the false negative error rate will need to be verified. Go to Step 5.

STEP 5: To calculate the power of the test, assume that the true values for the mean and standard deviation are those obtained in the sample and use a statistical software package like the DEFT software (EPA G-4D, 1994) or the DataQUEST software (EPA G-9D, 1996) to generate the power curve of the test.

If only one false negative error rate (β) has been specified (at P_1), it is possible to calculate the sample size which achieves the DQOs. To do this, calculate

$$m = \left[\frac{z_{1-\alpha}\sqrt{P_0(1-P_0)} + z_{1-\beta}\sqrt{P_1(1-P_1)}}{P_1 - P_0}\right]^2$$

If $m \le n$, the false negative error rate has been satisfied. Otherwise, the false negative error rate has not been satisfied.

STEP 6: The results of the test may be:

1) the null hypothesis was rejected and it seems that the proportion is greater than {less than} P_0;

2) the null hypothesis was not rejected, the false negative error rate was satisfied, and it seems that proportion is less than {greater than} P_0; or

3) the null hypothesis was not rejected, the false negative error rate was not satisfied, and it would seem the proportion was less than {greater than} P_0, but the conclusions are uncertain because the sample size was too small.

Box 3.2-9: An Example of the One-Sample Test for Proportions for a Simple Random Sample

Consider 85 samples of which 11 samples have concentrations greater than the clean-up standard. This data will be used to test the null hypothesis H_0: $P \geq .20$ vs. H_A: $P < .20$ (Case 2). The decision maker has specified a 5% false positive rate (α) for $P_0 = .2$, and a false negative rate (β) of 20% for $P_1 = 0.15$.

STEP 1: From the data, the observed proportion (p) is p = 11/85 = .1294

STEP 2: np = (85)(.1294) = 11 and n(1-p) = (85)(1-.1294) = 74. Since both np and n(1-p) are greater than or equal to 5, Steps 3 and 4 will be used.

STEP 3: Because H_0: $P \geq .20$, Case 2 formulas will be used.

$$z = \frac{p + .5/n - P_0}{\sqrt{P_0(1-P_0)/n}} = \frac{.1294 + .5/85 - .2}{\sqrt{.2(1-.2)/85}} = -1.492$$

STEP 4: Using Table A-1 of Appendix A, it was found that $z_{1-.05} = z_{.95} = 1.645$. Because $z \not< -z_{1-\alpha}$ (i.e., $-1.492 \not< -1.645$), the null hypothesis is not rejected so Step 5 will need to be completed.

STEP 5: To determine whether the test was powerful enough, the sample size necessary to achieve the DQOs was calculated as follows:

$$m = \left[\frac{1.64\sqrt{.2(1-.2)} + 1.04\sqrt{.15(1-.15)}}{.15 - .2}\right]^2 = 422.18$$

So 423 samples are required, many more than were actually taken.

STEP 6: The null hypothesis was not rejected and the false negative error rate was not satisfied. Therefore, it would seem the proportion is greater than 0.2, but this conclusion is uncertain because the sample size is too small.

3.2.3 Tests for a Median

A population median ($\tilde{\mu}$) is another measure of the center of the population distribution. This population parameter is less sensitive to extreme values and nondetects than the sample mean. Therefore, this parameter is sometimes used instead of the mean when the data contain a large number of nondetects or extreme values. The hypotheses considered in this section are:

Case 1: H_0: $\tilde{\mu} \leq C$ vs. H_A: $\tilde{\mu} > C$; and

Case 2: H_0: $\tilde{\mu} \geq C$ vs. H_A: $\tilde{\mu} < C$

where C represents a given threshold such as a regulatory level.

It is worth noting that the median is the 50 th percentile, so the methods described in section 3.2.2 may be used to test hypotheses concerning the median by letting $P_0 = 0.50$. In this case, the one-sample test for proportions is also called the Sign Test for a median. The Wilcoxon signed rank test (section 3.2.1.2) can also be applied to a median in the same manner as it is applied to a mean. In addition, this test is more powerful than the Sign Test for symmetric distributions. Therefore, the Wilcoxon signed rank test is the preferred test for the median.

3.3 TESTS FOR COMPARING TWO POPULATIONS

A two-sample test involves the comparison of two populations or a "before and after" comparison. In environmental applications, the two populations to be compared may be a potentially contaminated area with a background area or concentration levels from an upgradient and a downgradient well. The comparison of the two populations may be based on a statistical parameter that characterizes the relative location (e.g., a mean or median), or it may be based on a distribution-free comparison of the two population distributions. Tests that do not assume an underlying distributions (e.g., normal or lognormal) are called distribution-free or nonparametric tests. These tests are often more useful for comparing two populations than those that assume a specific distribution because they make less stringent assumptions. Section 3.3.1 covers tests for differences in the means of two populations. Section 3.3.2 covers tests for differences in the proportion or percentiles of two populations. Section 3.3.3 describes distribution-free comparisons of two populations. Section 3.3.4 describes tests for comparing two medians.

Often, a two-sample test involves the comparison of the difference of two population parameters to a threshold value. For environmental applications, the threshold value is often zero, representing the case where the data are used to determine which of the two population parameters is greater than the other. For example, concentration levels from a Superfund site may be compared to a background site. Then, if the Superfund site levels exceed the background levels, the site requires further investigation. A two-sample test may also be used to compare readings from two instruments or two separate populations of people.

If the exact same sampling locations are used for both populations, then the two samples are not independent. This case should be converted to a one-sample problem by applying the methods described in section 3.2 to the differences between the two populations at the same location. For example, one could compare contaminant levels from several wells after treatment to contaminant levels from the same wells before treatment. The methods described in section 3.2 would then be applied to the differences between the before and after treatment contaminant levels for each well.

3.3.1 Comparing Two Means

Let μ_1 represent the mean of population 1 and μ_2 represent the mean of population 2. The hypotheses considered in this section are:

Case 1: H_0: $\mu_1 - \mu_2 \leq \delta_0$ vs. H_A: $\mu_1 - \mu_2 > \delta_0$; and

Case 2: H_0: $\mu_1 - \mu_2 \geq \delta_0$ vs. H_A: $\mu_1 - \mu_2 < \delta_0$.

An example of a two-sample test for population means is comparing the mean contaminant level at a remediated Superfund site to a background site; in this case, δ_0 would be zero. Another example is a Record of Decision for a Superfund site which specifies that the remediation technique must reduce the mean contaminant level by 50 ppm each year. Here, each year would be considered a separate population and δ_0 would be 50 ppm.

The information required for these tests includes the null and alternative hypotheses (either Case 1 or Case 2); the gray region (i.e., a value $\delta_1 > \delta_0$ for Case 1 or a value $\delta_1 < \delta_0$ for Case 2 representing the bound of the gray region); the false positive error rate α at δ_0; the false negative error rate β at δ_1; and any additional limits on decision errors. It may be helpful to label additional false positive error limits as α_2 at $\delta_{\alpha 2}$, α_3 at $\delta_{\alpha 3}$, etc., and to label additional false negative error limits as β_2 at $\delta_{\beta 2}$, β_3 at $\delta_{\beta 3}$, etc.

3.3.1.1 Student's Two-Sample t-Test (Equal Variances)

PURPOSE

Student's two-sample t-test can be used to compare two population means based on the independent random samples $X_1, X_2, \ldots, X_m$ from the first population, and $Y_1, Y_2, \ldots, Y_n$ from the second population. This test assumes the variabilities (as expressed by the variance) of the two populations are approximately equal. If the two variances are not equal (a test is described in section 4.5), use Satterthwaite's t test (section 3.3.1.2).

ASSUMPTIONS AND THEIR VERIFICATION

The principal assumption required for the two-sample t-test is that a random sample of size m ($X_1, X_2, \ldots, X_m$) is drawn from population 1, and an independent random sample of size n ($Y_1, Y_2, \ldots, Y_n$) is drawn from population 2. Validity of the random sampling and independence assumptions should be confirmed by reviewing the procedures used to select the sampling points.

The second assumption required for the two-sample t-tests are that the sample means $\bar{X}$ (sample 1) and $\bar{Y}$ (sample 2) are approximately normally distributed. If both m and n are large, one may make this assumption without further verification. For small sample sizes, approximate normality of the sample means can be checked by testing the normality of each of the two samples.

LIMITATIONS AND ROBUSTNESS

The two-sample t-test with equal variances is robust to violations of the assumptions of normality and equality of variances. However, if the investigator has tested and rejected normality or equality of variances, then nonparametric procedures may be applied. The t-test is not robust to outliers because sample means and standard deviations are sensitive to outliers.

SEQUENCE OF STEPS

Directions for the two-sample t-test for a simple random sample and a systematic simple random sample are given in Box 3.3-1 and an example in Box 3.3-2.

3.3.1.2 Satterthwaite's Two-Sample t-Test (Unequal Variances)

Satterthwaite's t-test should be used to compare two population means when the variances of the two populations are not equal. It requires the same assumptions as the two-sample t-test (section 3.3.1.1) except the assumption of equal variances.

Directions for Satterthwaite's t-test for a simple random sample and a systematic simple random sample are given in Box 3.3-3 and an example in Box 3.3-4.

Box 3.3-1: Directions for the Student's Two-Sample t-Test (Equal Variances) for Simple and Systematic Random Samples

This describes the steps for applying the two-sample t-tests for differences between the population means when the two population variances are equal for Case 1 (H_0: $\mu_1 - \mu_2 \le \delta_0$). Modifications for Case 2 (H_0: $\mu_1 - \mu_2 \ge \delta_0$) are given in parentheses { }.

STEP 1: Calculate the sample mean $\overline{X}$ and the sample variance s_X^2 for sample 1 and compute the sample mean $\overline{Y}$ and the sample variance s_Y^2 for sample 2.

STEP 2: Use section 4.5 to determine if the variances of the two populations are equal. If the variances of the two populations are not equal, use Satterthwaite's t test (section 3.3.1.2). Otherwise, compute the pooled standard deviation

$$s_E = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{(m-1)+(n-1)}}.$$

STEP 3: Calculate $t = \dfrac{\overline{X} - \overline{Y} - \delta_0}{s_E\sqrt{1/n + 1/m}}$.

Use Table A-1 of Appendix A to find the critical value $t_{1-\alpha}$ such that 100(1-α)% of the t-distribution with (m+n-2) degrees of freedom is below $t_{1-\alpha}$.

If $t > t_{1-\alpha}$ {$t < -t_{1-\alpha}$}, the null hypothesis may be rejected. Go to Step 5.

If $t \not> t_{1-\alpha}$ {$t \not< -t_{1-\alpha}$}, there is not enough evidence to reject the null hypothesis. Therefore, the false negative error rate will need to be verified. Go to Step 4.

STEP 4: To calculate the power of the test, assume that the true values for the mean and standard deviation are those obtained in the sample and use a statistical software package like the DEFT software (EPA G-4D, 1994) or the DataQUEST software (EPA G-9D, 1996) to generate the power curve of the two-sample t-test. If only one false negative error rate (β) has been specified (at δ_1), it is possible to calculate the sample size which achieves the DQOs, assuming the true mean and standard deviation are equal to the values estimated from the sample, instead of calculating the power of the test. Calculate

$$m^* = n^* = \frac{2s^2(z_{1-\alpha}+z_{1-\beta})^2}{(\delta_1-\delta_0)^2} + (0.25)z_{1-\alpha}^2$$

If $m^* \le m$ and $n^* \le n$, the false negative error rate has been satisfied. Otherwise, the false negative error rate has not been satisfied.

STEP 5: The results of the test could be:

1) the null hypothesis was rejected, and it seems $\mu_1 - \mu_2 > \delta_0$ {$\mu_1 - \mu_2 < \delta_0$};

2) the null hypothesis was not rejected, the false negative error rate was satisfied, and it seems $\mu_1 - \mu_2 \le \delta_0$ {$\mu_1 - \mu_2 \ge \delta_0$}; or

3) the null hypothesis was not rejected, the false negative error rate was not satisfied, and it seems $\mu_1 - \mu_2 \le \delta_0$ {$\mu_1 - \mu_2 \ge \delta_0$}, but this conclusion is uncertain because the sample size was too small.

Box 3.3-2: An Example of a Student's Two-Sample t-Test (Equal Variances) for Simple and Systematic Random Samples

At a hazardous waste site, area 1 (cleaned using an in-situ methodology) was compared with a similar (but relatively uncontaminated) reference area, area 2. If the in-situ methodology worked, then the two sites should be approximately equal in average contaminant levels. If the methodology did not work, then area 1 should have a higher average than the reference area. Seven random samples were taken from area 1, and eight were taken from area 2. Because the contaminant concentrations in the two areas are supposedly equal, the null hypothesis is H_0: $\mu_1 - \mu_2 \leq 0$ (Case 1). The false positive error rate was set at 5% and the false negative error rate was set at 20% (β) if the difference between the areas is 2.5 ppb.

STEP 1:

	Sample Mean	Sample Variance
Area 1	7.8 ppm	2.1 ppm^2
Area 2	6.6 ppm	2.2 ppm^2

STEP 2: Methods described in section 4.5 were used to determine that the variances were essentially equal. Therefore,

$$s_E = \sqrt{\frac{(7-1)2.1 + (8-1)2.2}{(7-1)+(8-1)}} = 1.4676$$

STEP 3:

$$t = \frac{7.8 - 6.6 - 0}{1.4676\sqrt{1/7 + 1/8}} = 1.5798$$

Table A-1 of Appendix A was used to find that the critical value $t_{0.95}$ with (7 + 8 - 2) = 13 degrees of freedom is 1.771.

Because $t \not> t_{1-\alpha}$ (i.e., 1.5798 $\not>$ 1.771), there is not enough evidence to reject the null hypothesis. The false negative error rate will need to be verified.

STEP 4: Assuming the true values for the mean and standard deviation are those obtained in the sample:

$$m^* = n^* = \frac{2(1.4676^2)(1.645 + 0.842)^2}{(2.5-0)^2} + (0.25)1.645^2 = 4.938, \text{ i.e., } 5.$$

Because $m^* \leq m$ (7) and $n^* \leq n$ (8), the false negative error rate has been satisfied.

STEP 5: The null hypothesis was not rejected and the false negative error rate was satisfied. Therefore, it seems there is no difference between the two areas and that the in-situ methodology worked as expected.

Box 3.3-3: Directions for Satterthwaite's t-Test (Unequal Variances) for Simple and Systematic Random Samples

This describes the steps for applying the two-sample t-test for differences between the population means for Case 1 (H_0: $\mu_1 - \mu_2 \le \delta_0$). Modifications for Case 2 (H_0: $\mu_1 - \mu_2 \ge \delta_0$) are given in parentheses { }.

STEP 1: Calculate the sample mean $\overline{X}$ and the sample variance s_X^2 for sample 1 and compute the sample mean $\overline{Y}$ and the sample variance s_Y^2 for sample 2.

STEP 2: Using section 4.5, test whether the variances of the two populations are equal. If the variances of the two populations are not equal, compute: $s_{NE} = \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}$

If the variances of the two populations appear approximately equal, use Student's two-sample t-test (section 3.3.1.1, Box 3.3-1).

STEP 3: Calculate $t = \frac{\overline{X} - \overline{Y} - \delta_0}{s_{NE}}$.

Use Table A-1 of Appendix A to find the critical value $t_{1-\alpha}$ such that 100(1-α)% of the t-distribution with f degrees of freedom is below $t_{1-\alpha}$, where

$$f = \frac{\left[\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right]^2}{\frac{s_X^4}{m^2(m-1)} + \frac{s_Y^4}{n^2(n-1)}}$$

(Round f <u>down</u> to the nearest integer.)

If $t > t_{1-\alpha}$ {$t < -t_{1-\alpha}$}, the null hypothesis may be rejected. Go to Step 5.

If $t \ngtr t_{1-\alpha}$ {$t \nless -t_{1-\alpha}$}, there is not enough evidence to reject the null hypothesis and therefore, the false negative error rate will need to be verified. Go to Step 4.

STEP 4: If the null hypothesis (H_0) was not rejected, calculate either the power of the test or the sample size necessary to achieve the false positive and false negative error rates. To calculate the power of the test, assume that the true values for the mean and standard deviation are those obtained in the sample and use a statistical software package to generate the power curve of the two-sample t-test. A simple method to check on statistical power does not exist.

STEP 5: The results of the test could be:

1) the null hypothesis was rejected, and it seems $\mu_1 - \mu_2 > \delta_0$ {$\mu_1 - \mu_2 < \delta_0$};

2) the null hypothesis was not rejected, the false negative error rate was satisfied, and it seems $\mu_1 - \mu_2 \le \delta_0$ {$\mu_1 - \mu_2 \ge \delta_0$}; or

3) the null hypothesis was not rejected, the false negative error rate was not satisfied, and it seems $\mu_1 - \mu_2 \le \delta_0$ {$\mu_1 - \mu_2 \ge \delta_0$}, but this conclusion is uncertain because the sample size was too small.

Box 3.3-4: An Example of Satterthwaite's t-Test (Unequal Variances) for Simple and Systematic Random Samples

At a hazardous waste site, area 1 (cleaned using an in-situ methodology) was compared with a similar (but relatively uncontaminated) reference area, area 2. If the in-situ methodology worked, then the two sites should be approximately equal in average contaminant levels. If the methodology did not work, then area 1 should have a higher average than the reference area. Seven random samples were taken from area 1, and eight were taken from area 2. Because the contaminant concentrations in the two areas are supposedly equal, the null hypothesis is H_0: $\mu_1 - \mu_2 \leq 0$ (Case 1). The false positive error rate was set at 5% and the false negative error rate was set at 20% (β) if the difference between the areas is 2.5 ppb.

STEP 1:

	Sample Mean	Sample Variance
Area 1	9.2 ppm	1.3 ppm^2
Area 2	6.1 ppm	5.7 ppm^2

STEP 2: Using section 4.5, it was determined that the variances of the two populations were not equal, and therefore using Satterthwaite's method is appropriate:

$$s_{NE} = \sqrt{1.3/7 + 5.7/8} = 0.9477$$

STEP 3:

$$t = \frac{9.2 - 6.1 - 0}{0.9477} = 3.271$$

Table A-1 was used with f degrees of freedom, where

$$f = \frac{\left[1.3/7 + 5.7/8\right]^2}{\frac{1.3^2}{7^2(7-1)} + \frac{5.7^2}{8^2(8-1)}} = 10.307\ (i.e.,\ 10\ degrees\ of\ freedom)$$

(recall that f is rounded down to the nearest integer), to find $t_{1-\alpha} = 1.812$.

Because $t > t_{0.95}$ (3.271 > 1.812), the null hypothesis may be rejected.

STEP 5: Because the null hypothesis was rejected, it would appear there is a difference between the two areas (area 1 being more contaminated than area 2, the reference area) and that the in-situ methodology has not worked as intended.

3.3.2 Comparing Two Proportions or Percentiles

This section considers hypotheses concerning two population proportions (or two population percentiles); for example, one might use these tests to compare the proportion of children with elevated blood lead in one urban area compared with the proportion of children with elevated blood lead in another area. The population proportion is the ratio of the number of elements in a subset of the total population to the total number of elements, where the subset has some specific characteristic that the rest of the elements do not. A population percentile represents the percentage of elements of a population having values less than some threshold value C.

Let P_1 represent the true proportion for population 1, and P_2 represent the true proportion of population 2. The hypotheses considered in this section are:

Case 1: H_0: $P_1 - P_2 \leq \delta_0$ vs. H_A: $P_1 - P_2 > \delta_0$; and

Case 2: H_0: $P_1 - P_2 \geq \delta_0$ vs. H_A: $P_1 - P_2 < \delta_0$.

An equivalent null hypothesis for Case 1, written in terms of percentiles, is H_0: the $100P_1^{th}$ percentile minus the $100P_2^{th}$ percentile is C or larger, the reverse applying to Case 2. Since any hypothesis about the proportion below a threshold can be converted to an equivalent hypothesis about percentiles (see section 3.2.2), this guidance will only consider hypotheses concerning proportions.

The information required for this test includes the null and alternative hypotheses (either Case 1 or Case 2); the gray region (i.e., a value $\delta_1 > \delta_0$ for Case 1 or a value $\delta_1 < \delta_0$ for Case 2, representing the bound of the gray region); the false positive error rate α at δ_0; the false negative error rate β at δ_1; and any additional limits on decision errors.

3.3.2.1 Two-Sample Test for Proportions

PURPOSE

The two-sample test for proportions can be used to compare two population percentiles or proportions and is based on an independent random sample of m ($X_1, X_2, \ldots, X_m$) from the first population and an independent random sample size n ($Y_1, Y_2, \ldots, Y_n$) from the second population.

ASSUMPTIONS AND THEIR VERIFICATION

The principal assumption is that of random sampling from the two populations.

LIMITATIONS AND ROBUSTNESS

The two-sample test for proportions is valid (robust) for any underlying distributional shape and is robust to outliers, providing they are not pure data errors.

SEQUENCE OF STEPS

Directions for a two-sample test for proportions for a simple random sample and a systematic simple random sample are given in Box 3.3-5; an example is provided in Box 3.3-6.

Box 3.3-5: Directions for a Two-Sample Test for Proportions for Simple and Systematic Random Samples

The following describes the steps for applying the two-sample test for proportions for Case 1 (H_0: $P_1 - P_2 \le 0$). Modifications for Case 2 (H_0: $P_1 - P_2 \ge 0$) are given in braces { }.

STEP 1: Given m random samples $X_1, X_2, \ldots, X_m$ from the first population, and n samples from the second population, $Y_1, Y_2, \ldots, Y_n$, let k_1 be the number of points from sample 1 which exceed C, and let k_2 be the number of points from sample 2 which exceed C. Calculate the sample proportions $p_1 = k_1/m$ and $p_2 = k_2/n$. Then calculate the pooled proportion

$$p = (k_1 + k_2) / (m + n).$$

STEP 2: Compute mp_1, $m(1-p_1)$, np_2, $n(1-p_2)$. If all of these values are greater than or equal to 5, continue. Otherwise, seek assistance from a statistician as analysis is complicated.

STEP 3: Calculate $z = (p_1 - p_2) / \sqrt{p(1 - p)(1/m + 1/n)}.$

Use Table A-1 of Appendix A to find the critical value $z_{1-\alpha}$ such that 100(1-α)% of the normal distribution is below $z_{1-\alpha}$. For example, if $\alpha = 0.05$ then $z_{1-\alpha} = 1.645$.

If $z > z_{1-\alpha}$ {$z < -z_{1-\alpha}$}, the null hypothesis may be rejected. Go to Step 5.

If $z \not> z_{1-\alpha}$ {$z \not< -z_{1-\alpha}$}, there is not enough evidence to reject the null hypothesis. Therefore, the false negative error rate will need to be verified. Go to Step 4.

STEP 4: If the null hypothesis (H_0) was not rejected, calculate either the power of the test or the sample size necessary to achieve the false positive and false negative error rates. If only one false negative error rate (β) has been specified at $P_1 - P_2$, it is possible to calculate the sample sizes that achieve the DQOs (assuming the proportions are equal to the values estimated from the sample) instead of calculating the power of the test. To do this, calculate

$$m^* = n^* = \frac{2(z_{1-\alpha} + z_{1-\beta})^2 \bar{P}(1 - \bar{P})}{(P_2 - P_1)^2} \quad \text{where} \quad \bar{P} = \frac{P_1 + P_2}{2}.$$

and z_p is the pth percentile of the standard normal distribution (Table A-1 of Appendix A). If both m and n exceed m^*, the false negative error rate has been satisfied. If both m and n are below m^*, the false negative error rate has not been satisfied.

If m^* is between m and n, use a software package like the DEFT software (EPA G-4D, 1994) or the DataQUEST software (EPA G-9D, 1996) to calculate the power of the test, assuming that the true values for the proportions P_1 and P_2 are those obtained in the sample. If the estimated power is below 1-β, the false negative error rate has not been satisfied.

STEP 5: The results of the test could be:

1) the null hypothesis was rejected, and it seems the difference in proportions is greater than 0 {less than 0};

2) the null hypothesis was not rejected, the false negative error rate was satisfied, and it seems the difference in proportions is less than or equal to 0 {greater than or equal to 0}; or

3) the null hypothesis was not rejected, the false negative error rate was not satisfied, and it seems the difference in proportions is less than or equal to 0 {greater than or equal to 0}, but this outcome is uncertain because the sample size was probably too small.

Box 3.3-6: An Example of a Two-Sample Test for Proportions for Simple and Systematic Random Samples

At a hazardous waste site, investigators must determine whether an area suspected to be contaminated with dioxin needs to be remediated. The possibly contaminated area (area 1) will be compared to a reference area (area 2) to see if dioxin levels in area 1 are greater than dioxin levels in the reference area. An inexpensive surrogate probe was used to determine if each individual sample is either "contaminated," i.e., over the health standard of 1 ppb, or "clean," i.e., less than the health standard of 1 ppb. The null hypothesis will be that the proportion or contaminant levels in area 1 is less than or equal to the proportion in area 2, or H_0: $P_1 - P_2 \leq 0$ (Case 1). The decision maker is willing to accept a false positive decision error rate of 10% (α) and a false-negative decision error rate of 5% (β) when the difference in proportions between areas exceeds 0.10. A team collected 92 readings from area 1 (of which 12 were contaminated) and 80 from area 2, the reference area, (of which 10 were contaminated).

STEP 1: The sample proportion for area 1 is $p_1 = 12/92 = 0.130$, the sample proportion for area 2 is $p_2 = 10/80 = 0.125$, and the pooled proportion $p = (12 + 10) / (92 + 80) = 0.128$.

STEP 2: $mp_1 = 12$, $m(1-p_1) = 80$, $np_2 = 10$, $n(1-p_2) = 70$. Because these values are greater than or equal to 5, continue to step 3.

STEP 3: $$z = (0.130 - 0.125) / \sqrt{0.128(1 - 0.128)(1/92 + 1/80)} = 0.098$$

Table A-1 of Appendix A was used to find the critical value $z_{0.90} = 1.282$.

Because $z \not> z_{0.90}$ ($0.098 \not> 1.282$), there is not enough evidence to reject the null hypothesis and the false negative error rate will need to be verified. Go to Step 4.

STEP 4: Because the null hypothesis (H_0) was not rejected, calculate the sample size necessary to achieve the false positive and false negative error rates. Because only one false negative error rate ($\beta = 0.05$) has been specified (at a difference of $P_1 - P_2 = 0.1$), it is possible to calculate the sample sizes that achieve the DQOs, assuming the proportions are equal to the values estimated from the sample:

$$m^* = n^* = \frac{2(1.282 + 1.645)^2 0.1275(1 - 0.1275)}{(0.1)^2} = 190.6 \ (i.e., 191 \ samples)$$

where $$0.1275 = \bar{P} = \frac{0.115 + 0.055}{2}$$

Because both m and n are less than m^*, the false negative error rate has not been satisfied.

STEP 5: The null hypothesis was not rejected, and the false negative error rate was not satisfied. Therefore, it seems that there is no difference in proportions and that the contaminant concentrations of the investigated area and the reference area are probably the same. However, this outcome is uncertain because the sample sizes obtained were in all likelihood too small.

3.3.3 Nonparametric Comparisons of Two Populations

In many cases, assumptions on distributional characteristics are difficult to verify or difficult to satisfy for both populations. In this case, several distribution-free test procedures are available that compare the shape and location of the two distributions instead of a statistical parameter (such as a mean or median). The statistical tests described below test the null hypothesis "H_0: the distributions of population 1 and population 2 are identical (or, the site is not more contaminated than background)" versus the alternative hypothesis "H_A: part of the distribution of population 1 is located to the right of the distribution of population 2 (or the site is more contaminated than background)." Because of the structure of the hypothesis tests, the labeling of populations 1 and 2 is of importance. For most environmental applications, population 1 is the area of interest (i.e., the potentially contaminated area) and population 2 is the reference area.

There is no formal statistical parameter of interest in the hypotheses stated above. However, the concept of false positive and false negative error rates still applies.

3.3.3.1 The Wilcoxon Rank Sum Test

PURPOSE

The Wilcoxon rank sum test can be used to compare two population distributions based on m independent random samples $X_1, X_2, \ldots, X_m$ from the first population, and n independent random samples $Y_1, Y_2, \ldots, Y_n$ from the second population. When applied with the Quantile test (section 3.3.3.2), the combined tests are most powerful for detecting true differences between two population distributions.

ASSUMPTIONS AND THEIR VERIFICATION

The validity of the random sampling and independence assumptions should be verified by review of the procedures used to select the sampling points. The two underlying distributions are assumed to have the same shape and dispersion, so that one distribution differs by some fixed amount (or is increased by a constant) when compared to the other distribution. For large samples, to test whether both site distributions have approximately the same shape, one can create and compare histograms for the samples.

LIMITATIONS AND ROBUSTNESS

The Wilcoxon signed rank test may produce misleading results if many data values are the same. When values are the same, their relative ranks are the same, and this has the effect of diluting the statistical power of the Wilcoxon rank sum test. Estimated concentrations should be reported for data below the detection limit, even if these estimates are negative, because their relative magnitude to the rest of the data is of importance. An important advantage of the Wilcoxon rank sum test is its partial robustness to outliers, because the analysis is conducted in terms of rankings of the observations. This limits the influence of outliers because a given data point can be no more extreme than the first or last rank.

SEQUENCE OF STEPS

Directions and an example for the Wilcoxon rank sum test are given in Box 3.3-7 and Box 3.3-8. However, if a relatively large number of samples have been taken, it is more efficient in terms of statistical power to use a large sample approximation to the Wilcoxon rank sum test (Box 3.3-9) to obtain the critical values of W.

Box 3.3-7: Directions for the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples

Let $X_1, X_2, \ldots, X_m$ represent the m data points from population 1 and $Y_1, Y_2, \ldots, Y_n$ represent the n data points from population 2 where both m and n are less than or equal to 10. For this test, the null hypothesis will be that there is no difference between the two populations. The alternative hypothesis will be that population 1 is located to the right of population 2 for Case 1 or that population 2 is located to the right of population 1 for Case 2. If either m or n is larger than 10, use Box 3.3-9.

STEP 1: List and rank the measurements from both populations from smallest to largest, keeping track of which population contributed each measurement. The rank of 1 is assigned to the smallest value, the rank of 2 to the second smallest value, and so forth. If there are ties, assign the average of the ranks that would otherwise have been assigned to the tied observations.

STEP 2: For Case 1, calculate W as the sum of the ranks of the data from population 2.
For Case 2, calculate W as the sum of the ranks of the data from population 1.

STEP 3: Calculate $W_{XY} = W - \frac{n(n+1)}{2}$ for Case 1 or calculate $W_{XY} = W - \frac{m(m+1)}{2}$ for Case 2.

STEP 4: Use Table A-7 of Appendix A to find the critical value w_α.

If $W_{XY} \leq w_\alpha$, the null hypothesis may be rejected. Go to Step 6.

If $W_{XY} > w_\alpha$, there is not enough evidence to reject the null hypothesis. Therefore, the false negative error rate will need to be verified. Go to Step 5.

STEP 5: If the null hypothesis (H_0) was not rejected, calculate either the power of the test or the sample size necessary to achieve the false positive and false negative error rates using a software package like the DEFT software (EPA G-4D, 1994) or the DataQUEST software (EPA G-9D, 1996). (Power calculations tend to be much more difficult for nonparametric procedures than for parametric procedures.) If only one false negative error rate (β) has been specified (at δ_1), it is possible to calculate the sample size that achieves the DQOs, assuming the true mean and standard deviation are equal to the values estimated from the sample, instead of calculating the power of the test. If m and n are large, calculate:

$$m^* = n^* = \frac{2s^2(z_{1-\alpha}+z_{1-\beta})^2}{(\delta_1-\delta_0)^2} + (0.25)z_{1-\alpha}^2$$

where z_p is the pth percentile of the standard normal distribution (Table A-1 of Appendix A). Then, multiply m* and n* by 1.16 to account for loss in efficiency, and, if $1.16m^* \leq m$ and $1.16n^* \leq n$, the false negative error rate has been satisfied; if the values of m and n are otherwise, the false negative error rate has not been satisfied.

STEP 6: The results of the test could be:

1) the null hypothesis was rejected, and it seems that population 1 is located to the right of population 2 for Case 1 or that population 2 is located to the right of population 1 for Case 2.

2) the null hypothesis was not rejected, the false negative error rate was satisfied, and it seems there is no difference between the two populations; or

3) the null hypothesis was not rejected, the false negative error rate was not satisfied, and it seems there is no difference between the two populations, but this result is uncertain because the sample sizes were probably too small.

Box 3.3-8: An Example of the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples

At a hazardous waste site, area 1 (cleaned using an in-situ methodology) was compared with a similar (but relatively uncontaminated) reference area, area 2. If the in-situ methodology worked, then the two sites should be approximately equal in average contaminant levels. If the methodology did not work, then area 1 should have a higher average than the reference area. The null hypothesis will be that there is no difference between the two areas. Since area 1 was previously contaminated, the alternative hypothesis will be that contaminant levels in area 1 are larger (located to the right) than those in area 2 (Case 1). The false positive error rate was set at 5% and the false negative error rate was set at 20% (β) if the difference between the areas is 2.5 ppb. Seven random samples were taken from area 1 and eight samples were taken from area 2:

Area 1	Area 2
17, 23, 26, 5	16, 20, 5, 4
13, 13, 12	8, 10, 7, 3

STEP 1: The data listed and ranked by size are (Area 1 denoted by *):

Data (ppb):	3,	4,	5,	5*,	7,	8*,	10,	12,	13*,	13*,	16,	17*,	20,	23*,	26*
Rank:	1,	2,	3.5,	3.5*,	5,	6*,	7,	8,	9.5*,	9.5*	11,	12*,	13,	14*,	15*

STEP 2: W = sum of ranks from area 2 = 50.5

STEP 3: $W_{XY} = 50.5 - 8(8 + 1)/2 = 14.5$

STEP 4: Using Table A-7 of Appendix A, $W_{0.05} = 13$. Because W_{XY} is greater than $W_{0.05}$, do not reject the null hypothesis.

STEP 5: The null hypothesis was not rejected and it would be appropriate to calculate the probable power of the test. However, because the number of samples is small, extensive computer simulations are required in order to estimate the power of this test. Therefore, a statistician should be consulted.

STEP 6: The null hypothesis was not rejected. Therefore, it is likely that there is no difference between the investigated area and the reference area, although the statistical power is low due to the small sample sizes involved.

Box 3.3-9: Directions for the Large Sample Approximation to the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples

Let $X_1, X_2, \ldots, X_m$ represent the m data points from population 1 and $Y_1, Y_2, \ldots, Y_n$ represent the n data points from population 2 where both n and m are greater than 10. The null hypothesis will be that there is no difference between the two populations. The alternative hypothesis will be that population 1 is larger than population 2 for Case 1 or that population 2 is larger than population 1 for Case 2.

STEP 1: List and rank the measurements from both populations from smallest to largest, keeping track of which population contributed each measurement. The rank of 1 is assigned to the smallest value, the rank of 2 to the second smallest value, and so forth. If there are ties, assign the average of the ranks that would otherwise have been assigned to the tied observations.

STEP 2: For Case 1, calculate W as the sum of the ranks of the data from population 2.
For Case 2, calculate W as the sum of the ranks of the data from population 1.

STEP 3: Calculate $z = \dfrac{W - \dfrac{n(m+n+1)}{2}}{\sqrt{mn(m+n+1)/2}}$ for Case 1 or $z = \dfrac{W - \dfrac{m(m+n+1)}{2}}{\sqrt{mn(m+n+1)/2}}$ for Case 2.

STEP 4: Use Table A-1 of Appendix A to find the critical value $z_{1-\alpha}$ such that 100(1-α)% of the normal distribution is below $z_{1-\alpha}$.

If $z > z_{1-\alpha}$, there is not enough evidence to reject the null hypothesis and the false negative error rate should be verified. Go to Step 5.

If $z \leq z_{1-\alpha}$, the null hypothesis may be rejected. Go to Step 6.

STEP 4: If the null hypothesis (H_0) was not rejected, calculate either the power of the test or the sample size necessary to achieve the false positive and false negative error rates using a statistical software package. (Power calculations tend to be more difficult for nonparametric procedures than for parametric procedures.) If only one false negative error rate (β) has been specified (at δ_1), it is possible to calculate the sample size that achieves the DQOs, assuming the true mean and standard deviation are equal to the values estimated from the sample, instead of calculating the power of the test. If m and n are large, calculate:

$$m^* = n^* = \frac{2s^2(z_{1-\alpha}+z_{1-\beta})^2}{(\delta_1-\delta_0)^2} + (0.25)z_{1-\alpha}^2$$

where z_p is the pth percentile of the standard normal distribution (Table A-1 of Appendix A). Then, multiply m* and n* by 1.16 to account for a loss in efficiency. If $1.16m^* \leq m$ and $1.16n^* \leq n$, the false negative error rate has been satisfied. Otherwise, the false negative error rate has not been satisfied.

STEP 6: The results of the test could be:

1) the null hypothesis was rejected, and it seems that population 1 is greater than population 2 for Case 1 or that population 2 is greater than population 1 for Case 2.

2) the null hypothesis was not rejected, the false negative error rate was satisfied, and it seems there is no difference between the two populations; or

3) the null hypothesis was not rejected, the false negative error rate was not satisfied, and it seems there is no difference between the two populations, but this result is uncertain because the sample sizes were probably too small.

3.3.3.2 The Quantile Test

PURPOSE

The Quantile test can be used to compare two populations based on the independent random samples $X_1, X_2, \ldots, X_m$ from the first population and $Y_1, Y_2, \ldots, Y_n$ from the second population. When the Quantile test and the Wilcoxon rank sum test (section 3.3.3.1) are applied together, the combined tests are the most powerful at detecting true differences between two populations.

ASSUMPTIONS AND THEIR VERIFICATION

The Quantile test assumes that the data $X_1, X_2, \ldots, X_m$ are a random sample from population 1, and the data $Y_1, Y_2, \ldots, Y_n$ are a random sample from population 2, and the two random samples are independent of one another. The validity of the random sampling and independence assumptions is assured by using proper randomization procedures, either random number generators or tables of random numbers. The primary verification required is to review the procedures used to select the sampling points. The two underlying distributions are assumed to have the same underlying dispersion (variance).

LIMITATIONS AND ROBUSTNESS

The Quantile test is not robust to outliers. In addition, the test assumes either a systematic (e.g., a triangular grid) or simple random sampling was employed. The Quantile test may not be used for stratified designs.

SEQUENCE OF STEPS

The Quantile test is difficult to implement by hand. Therefore, directions are not included in this guidance. However, the DataQUEST software (EPA G-9D, 1996) can be used to conduct this test.

3.3.4 Comparing Two Medians

Let $\tilde{\mu}_1$ represent the median of population 1 and $\tilde{\mu}_2$ represent the median of population 2. The hypothesis considered in this section are:

Case 1: H_0: $\tilde{\mu}_1 - \tilde{\mu}_2 \leq \delta_0$ vs. H_A: $\tilde{\mu}_1 - \tilde{\mu}_2 > \delta_0$; and

Case 2: H_0: $\tilde{\mu}_1 - \tilde{\mu}_2 \geq \delta_0$ vs. H_A: $\tilde{\mu}_1 - \tilde{\mu}_2 < \delta_0$.

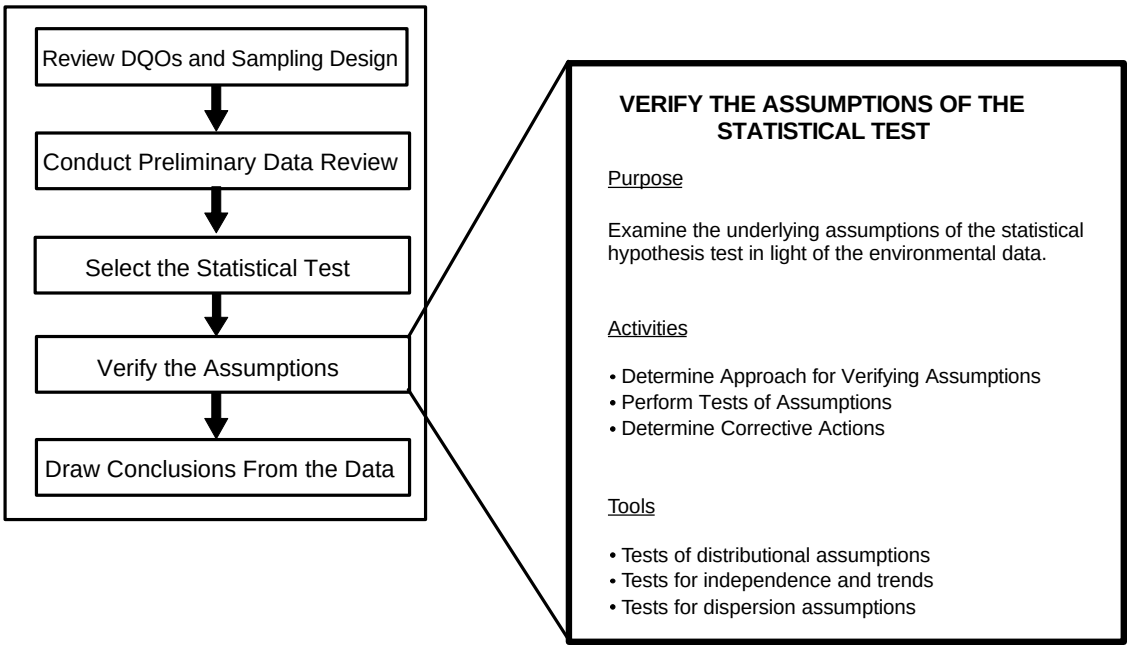
An example of a two-sample test for the difference between two population medians is comparing the median contaminant level at a Superfund site to the median of a background site. In this case, δ_0 would be zero.

The median is also the 50 th percentile, and, therefore, the methods described in section 3.3.2 for percentiles and proportions may be used to test hypotheses concerning the difference between two medians by letting $P_1 = P_0 = 0.50$. The Wilcoxon rank sum test (section 3.3.3.1) is also recommended for comparing two medians. This test is more powerful than those for proportions for symmetric distributions.

CHAPTER 4

STEP 4: VERIFY THE ASSUMPTIONS OF THE STATISTICAL TEST

THE DATA QUALITY ASSESSMENT PROCESS



Step 4: Verify the Assumptions of the Statistical Test

- Determine approach for verifying assumptions.
  - Identify any strong graphical evidence from the preliminary data review.
  - Review (or develop) the statistical model for the data.
  - Select the tests for verifying assumptions.
- Perform tests of assumptions.
  - Adjust for bias if warranted.
  - Perform the calculations required for the tests.
- If necessary, determine corrective actions.
  - Determine whether data transformations will correct the problem.
  - If data are missing, explore the feasibility of using theoretical justification or collecting new data.
  - Consider robust procedures or nonparametric hypothesis tests.

STEP 4: VERIFY THE ASSUMPTIONS OF THE STATISTICAL TEST

	Test	Section	Directions	Example
Tests for Distributional Assumptions	Shapiro Wilk W Test	4.2.2		
	Filliben's Statistic	4.2.3		
	Coefficient of Variation Test	4.2.4	Box 4.2-1	Box 4.2-1
	Skewness and Kurtosis Tests	4.2.5		
	Studentized Range Test	4.2.6	Box 4.2-2	Box 4.2-2
	Geary's Test	4.2.6	Box 4.2-3	Box 4.2-4
	Goodness-of-Fit Tests	4.2.7		
Tests for Trends	Test of a Correlation Coefficient	4.3.2.2	Box 4.3.1	Box 4.3.1
	Mann-Kendall Test	4.3.4.1 4.3.4.2	Box 4.3.3 Box 4.3.5	Box 4.3.4 Box 4.3.6
	Tests for an Overall Monotonic Trend	4.3.4.3	Box 4.3-8	
Tests for Outliers	Extreme Value Test	4.4.3	Box 4.4-1	Box 4.4-2
	Discordance Test	4.4.4	Box 4.4-3	Box 4.4-4
	Rosner's Test	4.4.5	Box 4.4-5	Box 4.4-6
	Walsh's Test	4.4.6	Box 4.4-7	
Tests for Dispersion	Confidence Intervals for a Variance	4.5.1	Box 4.5-1	Box 4.5-1
	F-Test	4.5.2	Box 4.5-2	Box 4.5-2
	Bartlett’s Test	4.5.3	Box 4.5-3	Box 4.5-4
	Levene’s Test	4.5.4	Box 4.5-5	Box 4.5-6
Transformations	*Logarithmic, Square Root, Inverse Sine, Box-Cox Transformations*	4.6	Box 4.6-1	Box 4.6-1
Data below Detection Limit	Substitution Methods	4.7.1		
	Cohen's Adjustment	4.7.2.1	Box 4.7-1	Box 4.7-2
	Trimmed Mean	4.7.2.2	Box 4.7-4	Box 4.7-5
	Winsorization	4.7.2.3	Box 4.7-6	Box 4.7-7

CHAPTER 4
STEP 4: VERIFY THE ASSUMPTIONS OF THE STATISTICAL TEST

4.1 OVERVIEW AND ACTIVITIES

In this step, the analyst should assess the validity of the statistical test chosen in step 3 by examining its underlying assumptions in light of the newly generated environmental data. The principal thrust of this section is the determination of whether the data support the underlying assumptions necessary for the selected test, or if modifications to the data are necessary prior to further statistical analysis.

This determination can be performed quantitatively using statistical analysis of data to confirm or reject the assumptions that accompany any statistical test. Almost always, however, the quantitative techniques must be supported by qualitative judgments based on the underlying science and engineering aspects of the study. Graphical representations of the data, such as those described in Chapter 2, can provide important qualitative information about the reasonableness of the assumptions. Documentation of this step is important, especially when subjective judgments play a pivotal role in accepting the results of the analysis.

If the data support all of the key assumptions of the statistical test, then the DQA Process continues to the next step, drawing conclusions from the data (Chapter 5). However, often one or more of the assumptions will be called into question which may trigger a reevaluation of one of the previous steps. This iteration in the DQA Process is an important check on the validity and practicality of the results.

4.1.1 Determine Approach for Verifying Assumptions

In most cases, assumptions about distributional form, independence, and dispersion can be verified formally using the statistical tests described in the technical sections in the remainder of this chapter, although in some situations, information from the preliminary data review may serve as sufficiently strong evidence to support the assumptions. As part of this activity, the analyst should identify methods to verify that the type and quantity of data required to perform the desired test are available. The outputs of this activity should include a list of the specific tests that will be used to verify the assumptions.

The methods and approach chosen for assumption verification depend on the nature of the study and its documentation. For example, if computer simulation was used to estimate the theoretical power of the statistical test, then this simulation model should be the basis for evaluation of the effect of changes to assumptions using estimates calculated from the data to replace simulation values.

If it is not already part of the design documentation, the analyst may need to formulate a statistical model that describes the data. In a statistical model, the data are conceptually decomposed into elements that are assumed to be “fixed” (i.e., the component is either a constant but unknown feature of the population or is controlled by experimentation) or “random” (i.e., the component is an uncontrolled source of variation). Which components are considered fixed and which are random is determined by the assumptions made for the statistical test and by the inherent structure of the sampling design. The random components that represent the sources of uncontrolled variation could include several types of measurement errors, as well as other sources such as temporal and/or spatial components.

In addition to identifying the components that make up an observation and specifying which are fixed and which are random, the model should also define whether the various components behave in an additive or

multiplicative fashion (or some combination). For example, if temporal or spatial autocorrelations are believed to be present, then the model needs to identify the autocorrelation structure (see section 2.3.8).

4.1.2 Perform Tests of Assumptions

For most statistical tests, investigators will need to assess the reasonableness of assumptions in relation to the structure of the components making up an observation. For example, a t-test assumes that the components, or errors, are additive, uncorrelated, and normally distributed with homogeneous variance. Basic assumptions that should be investigated include:

(1) *Is it reasonable to assume that the errors (deviations from the model) are normally distributed?* If adequate data are available, then standard tests for normality can be conducted (e.g., the Shapiro-Wilk test or the Kolmogorov-Smirnov test).

(2) *Is it reasonable to assume that errors are uncorrelated?* While it is natural to assume that analytical errors imbedded in measurements made on different sample units are independent, other errors from other sources may not be independent. If sample units are "too close together," either in time or space, independence may not hold. If the statistical test assumes independence and this assumption is not correct, the proposed false positive and false negative error rates (α and β) for the statistical test cannot be verified.

(3) *Is it reasonable to assume that errors are additive and have a constant variability?* If sufficient data are available, a plot of the relevant standard deviations versus mean concentrations may be used to discern if variability tends to increase with concentration level. If so, transformations of the data may make the additivity assumption more tenable.

One of the most important assumptions underlying the statistical procedures described herein is that there is no inherent bias (systematic deviation from the true value) in the data. The general approach adopted here is that if a long term bias is known to exist, then adjustment for this bias should be made. If bias is present, then the basic effect is to shift the power curves associated with a given test to the right or left, depending on the direction of the bias. Thus substantial distortion of the nominal Type I (false positive) and Type II (false negative) decision error rates may occur. In general, bias cannot be discerned by examination of routine data; rather, appropriate and adequate QA data are needed, such as performance evaluation data. If one chooses not to make adjustment for bias on the basis of such data, then one should, at a minimum, construct the estimated worse-case power curves so as to understand the potential effects of the bias.

4.1.3 Determine Corrective Actions

Sometimes the assumptions underlying the primary statistical test will not be satisfied and some type of corrective action will be required before proceeding. In some cases, a transformation of the data will correct a problem with distributional assumptions. In other cases, the data for verifying some key assumption may not be available, and existing information may not support a theoretical justification of the validity of the assumption. In this situation, it may be necessary to collect additional data to verify the assumptions. If the assumptions underlying a hypothesis test are not satisfied, and data transformations or other modifications do not appear feasible, then it may be necessary to consider an alternative statistical test. These include robust test procedures and nonparametric procedures. Robust test procedures involve modifying the parametric test by using robust estimators. For instance, as a substitute for a t-test, a trimmed mean and its associated standard error (section 4.7.2) might be used to form a t-type statistic.

4.2 TESTS FOR DISTRIBUTIONAL ASSUMPTIONS

Many statistical tests and models are only appropriate for data that follow a particular distribution. This section will aid in determining if a distributional assumption of a statistical test is satisfied, in particular, the assumption of normality. Two of the most important distributions for tests involving environmental data are the normal distribution and the lognormal distribution, both of which are discussed in this section. To test if the data follow a distribution other than the normal distribution or the lognormal distribution, apply the chi-square test discussed in section 4.2.7 or consult a statistician.

There are many methods available for verifying the assumption of normality ranging from simple to complex. This section discusses methods based on graphs, sample moments (kurtosis and skewness), sample ranges, the Shapiro-Wilk test and closely related tests, and goodness-of-fit tests. Discussions for the simplest tests contain step-by-step directions and examples based on the data in Table 4.2-1. These tests are summarized in Table 4.2-2. This section ends with a comparison of the tests to help the analyst select a test for normality.

Table 4.2-1. Data for Examples in Section 4.2

15.63	11.00	11.75	10.45	13.18	10.37	10.54	11.55	11.01	10.23	$\bar{x}$ =11.57 s = 1.677

The assumption of normality is very important as it is the basis for the majority of statistical tests. A normal, or Gaussian, distribution is one of the most common probability distributions in the analysis of environmental data. A normal distribution is a reasonable model of the behavior of certain random phenomena and can often be used to approximate other probability distributions. In addition, the Central Limit Theorem and other limit theorems state that as the sample size gets large, some of the sample summary statistics (e.g., the sample mean) behave as if they are a normally distributed variable. As a result, a common assumption associated with parametric tests or statistical models is that the errors associated with data or a model follow a normal distribution.

The graph of a normally distributed random variable, a normal curve, is bell-shaped (see Figure 4.2-1) with the highest point located at the mean which is equal to the median. A normal curve is symmetric

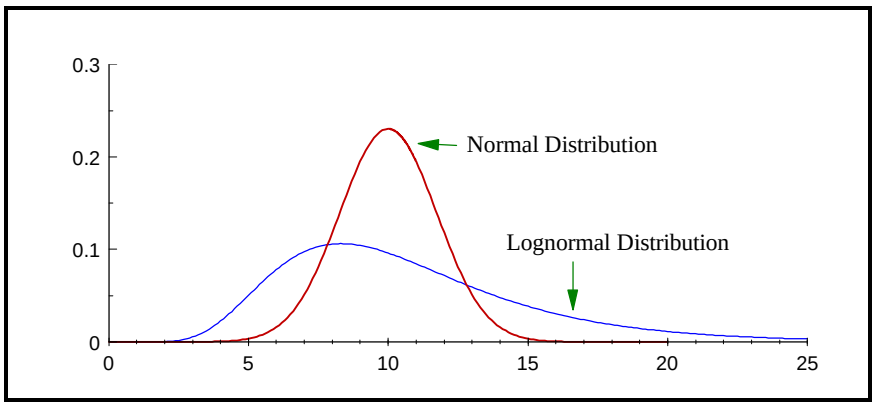


Figure 4.2-1. Graph of a Normal and Lognormal Distribution

about the mean, hence the part to the left of the mean is a mirror image of the part to the right. In environmental data, random errors occurring during the measurement process may be normally distributed.

Environmental data commonly exhibit frequency distributions that are non-negative and skewed with heavy or long right tails. Several standard parametric probability models have these properties, including the Weibull, gamma, and lognormal distributions. The lognormal distribution (Figure 4.2-1) is a commonly used distribution for modeling environmental contaminant data. The advantage to this distribution is that a simple (logarithmic) transformation will transform a lognormal distribution into a normal distribution. Therefore, the methods for testing for normality described in this section can be used to test for lognormality if a logarithmic transformation has been used.

Table 4.2-2. Tests for Normality

Test	Section	Sample Size	Recommended Use	Data-QUEST
Shapiro Wilk W Test	4.2.2	$\leq$ 50	Highly recommended.	Yes
Filliben's Statistic	4.2.3	$\leq$ 100	Highly recommended.	Yes
Coefficient of Variation Test	4.2.4	Any	Only use to quickly discard an assumption of normality.	Yes
Skewness and Kurtosis Tests	4.2.5	> 50	Useful for large sample sizes.	Yes
Geary's Test	4.2.6	> 50	Useful when tables for other tests are not available.	Yes
Studentized Range Test	4.2.6	$\leq$ 1000	Highly recommended (with some conditions).	Yes
Chi-Square Test	4.2.7	Large[a]	Useful for grouped data and when the comparison distribution is known.	No
Lilliefors Kolmogorov-Smirnoff Test	4.2.7	> 50	Useful when tables for other tests are not available.	No

[a] The necessary sample size depends on the number of groups formed when implementing this test. Each group should contain at least 5 observations.

4.2.1 Graphical Methods

Graphical methods (section 2.3) present detailed information about data sets that may not be apparent from a test statistic. Histograms, stem-and-leaf plots, and normal probability plots are some graphical methods that are useful for determining whether or not data follow a normal curve. Both the histogram and stem-and-leaf plot of a normal distribution are bell-shaped. The normal probability plot of a normal distribution follows a straight line. For non-normally distributed data, there will be large deviations in the tails or middle of a normal probability plot.

Using a plot to decide if the data are normally distributed involves making a subjective decision. For extremely non-normal data, it is easy to make this determination; however, in many cases the decision is not straightforward. Therefore, formal test procedures are usually necessary to test the assumption of normality.

4.2.2 Shapiro-Wilk Test for Normality (the W test)

One of the most powerful tests for normality is the W test by Shapiro and Wilk. This test is similar to computing a correlation between the quantiles of the standard normal distribution and the ordered values of a data set. If the normal probability plot is approximately linear (i.e., the data follow a normal curve), the test statistic will be relatively high. If the normal probability plot contains significant curves, the test statistic will be relatively low.

The W test is recommended in several EPA guidance documents and in many statistical texts. Tables of critical values for sample sizes up to 50 have been developed for determining the significance of the test statistic. However, this test is difficult to compute by hand since it requires two different sets of tabled values and a large number of summations and multiplications. Therefore, directions for implementing this test are not given in this document, but the test is contained in the DataQUEST software package (QA/G-9D, 1996).

4.2.3 Extensions of the Shapiro-Wilk Test (Filliben's Statistic)

Because the W test may only be used for sample sizes less than or equal to 50, several related tests have been proposed. D'Agostino’s test for sample sizes between 50 and 1000 and Royston’s test for sample sizes up to 2000 are two such tests that approximate some of the key quantities or parameters of the W test.

Another test related to the W test is the Filliben statistic, also called the probability plot correlation coefficient. This test measures the linearity of the points on the normal probability plot. Similar to the W test, if the normal probability plot is approximately linear (i.e., the data follow a normal curve), the correlation coefficient will be relatively high. If the normal probability plot contains significant curves (i.e., the data do not follow a normal curve), the correlation coefficient will be relatively low. Although easier to compute that the W test, the Filliben statistic is still difficult to compute by hand. Therefore, directions for implementing this test are not given in this guidance; however, it is contained in the DQA DataQUEST software package (QA/G-9D, 1996).

4.2.4 Coefficient of Variation

The coefficient of variation (CV) may be used to quickly determine whether or not the data follow a normal curve by comparing the sample CV to 1. The use of the CV is only valid for some environmental applications if the data represent a non-negative characteristic such as contaminant concentrations. If the CV is greater than 1, the data should not be modeled with a normal curve. However, this method *should not be used to conclude the opposite*, i.e., do not conclude that the data can be modeled with a normal curve if the CV is less than 1. This test is to be used only in conjunction with other statistical tests or when graphical representations of the data indicate extreme departures from normality. Directions and an example of this method are contained in Box 4.2-1.

4.2.5 Coefficient of Skewness/Coefficient of Kurtosis Tests

The degree of symmetry (or asymmetry) displayed by a data set is measured by the coefficient of skewness (g_3). The coefficient of kurtosis, g_4, measures the degree of flatness of a probability distribution near its center. Several test methods have been proposed using these coefficients to test for normality. One method tests for normality by adjusting the coefficients of skewness and kurtosis to approximate a standard normal distribution for sample sizes greater than 50.

Two other tests based on these coefficients include a combined test based on a chi-squared (χ^2) distribution and Fisher's cumulant test. Fisher's cumulant test computes the exact sampling distribution of g_3 and g_4; therefore, it is more powerful than previous methods which assume that the distributions of the two coefficients are normal. Fisher's cumulant test requires a table of critical values, and these tests require a sample size of greater than 50. Tests based on skewness and kurtosis are rarely used as they are difficult to compute and less powerful than many alternatives.

Box 4.2-1: Directions for the Coefficient of Variation Test for Environmental Data and an Example

Directions

STEP 1: Calculate the coefficient of variation (CV): $$CV = s / \bar{X} = \frac{\left[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right]^{1/2}}{\frac{1}{n} \sum_{i=1}^{n} X_i}$$

STEP 2: If CV > 1.0, conclude that the data are not normally distributed. Otherwise, the test is inconclusive.

Example

The following example demonstrates using the coefficient of variation to determine that the data in Table 4.2-1 should not be modeled using a normal curve.

STEP 1: Calculate the coefficient of variation (CV): $$CV = \frac{s}{\bar{X}} = \frac{1.677}{11.571} = 0.145$$

STEP 2: Since 0.145 $\not>$ 1.0, the test is inconclusive.

4.2.6 Range Tests

Almost 100% of the area of a normal curve lies within ±5 standard deviations from the mean and tests for normality have been developed based on this fact. Two such tests, which are both simple to apply, are the studentized range test and Geary's test. Both of these tests use a ratio of an estimate of the sample range to the sample standard deviation. Very large and very small values of the ratio then imply that the data are not well modeled by a normal curve.

a. The studentized range test (or w/s test). This test compares the range of the sample to the sample standard deviation. Tables of critical values for sample sizes up to 1000 (Table A-2 of Appendix A) are available for determining whether the absolute value of this ratio is significantly large. Directions for implementing this method are given in Box 4.2-2 along with an example. The studentized range test does not perform well if the data are asymmetric and if the tails of the data are heavier than the normal distribution. In addition, this test may be sensitive to extreme values. Unfortunately, lognormally distributed data, which are common in environmental applications, have these characteristics. If the data appear to be lognormally distributed, then this test should not be used. In most cases, the studentized range test performs as well as the Shapiro-Wilk test and is much easier to apply.

b. Geary's Test. Geary's test uses the ratio of the mean deviation of the sample to the sample standard deviation. This ratio is then adjusted to approximate a standard normal distribution. Directions for implementing this method are given in Box 4.2-3 and an example is given in Box 4.2-4. This test does not perform as well as the Shapiro-Wilk test or the studentized range test. However, since Geary's test statistic is based on the normal distribution, critical values for all possible sample sizes are available.

Box 4.2-2: Directions for Studentized Range Test and an Example

Directions

STEP 1: Calculate sample range (w) and sample standard deviation (s) using section 2.2.3.

STEP 2: Compare $\frac{w}{s} = \frac{X_{(n)} - X_{(1)}}{s}$ to the critical values given in Table A-2 (labeled a and b).

If w/s falls outside the two critical values then the data do not follow a normal curve.

Example

The following example demonstrates the use of the studentized range test to determine if the data from Table 4.2-1 can be modeled using a normal curve.

STEP 1: $w = X_{(n)} - X_{(1)} = 15.63 - 10.23 = 5.40$ and $s = 1.677$.

STEP 2: w/s = 5.4 / 1.677 = 3.22. The critical values given in Table A-2 are 2.51 and 3.875. Since 3.22 falls between these values, the assumption of normality is not rejected.

Box 4.2-3: Directions for Geary's Test

STEP 1: Calculate the sample mean $\bar{x}$, the sample sum of squares (SSS), and the sum of absolute deviations (SAD):

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad SSS = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}, \quad and \quad SAD = \sum_{i=1}^{n} |X_i - \bar{X}|$$

STEP 2: Calculate Geary's test statistic $a = \dfrac{SAD}{\sqrt{n(SSS)}}$

STEP 3: Test "a" for significance by computing $Z = \dfrac{a - 0.7979}{0.2123/\sqrt{n}}$. Here 0.7979 and 0.2123 are constants used to achieve normality.

STEP 4: Use Table A-1 of Appendix A to find the critical value $z_{1-\alpha}$ such that 100(1-α)% of the normal distribution is below $z_{1-\alpha}$. For example, if $\alpha = 0.05$, then $z_{1-\alpha} = 1.645$. Declare "a" to be sufficiently small or large (i.e., conclude the data are not normally distributed) if $|Z| > Z_{1-\alpha}$.

Box 4.2-4: Example of Geary's Test

The following example demonstrates the use of Geary's test to determine if the data from Table 4.2-1 can be modeled using a normal curve.

STEP 1: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 11.571, \quad SAD = \sum_{i=1}^{n} |X_i - \bar{X}| = 11.694, \; and$

$$SSS = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} = 1364.178 - 1338.88 = 25.298$$

STEP 2: $a = \dfrac{SAD}{\sqrt{n(SSS)}} = \dfrac{11.694}{\sqrt{10(25.298)}} = 0.735$

STEP 3: $Z = \dfrac{0.735 - 0.7979}{0.2123/\sqrt{10}} = -0.934$

STEP 4: Since $|Z| \not> 1.64$ (5% significance level), there is not enough information to conclude that the data do not follow a normal distribution.

4.2.7 Goodness-of-Fit Tests

Goodness-of-fit tests are used to test whether data follow a specific distribution, i.e., how “good” a specified distribution fits the data. In verifying assumptions of normality, one would compare the data to a normal distribution with a specified mean and variance.

a. Chi-square Test. One classic goodness-of-fit test is the chi-square test which involves breaking the data into groups and comparing these groups to the expected groups from the known distribution. There are no fixed methods for selecting these groups and this test also requires a large sample size since at least 5 observations per group are required to implement this test. In addition, the chi-square test does not have the power of the Shapiro-Wilk test or some of the other tests mentioned above.

b. Kolmogorov-Smirnoff (K-S) Test and Lilliefors K-S Test. Another goodness-of-fit test is the Kolmogorov Smirnoff (K-S) test which also tests whether the data follow a specific distribution with known parameters such as the mean and variance. This test requires that the sample size of the data be greater than 50. The Lilliefors K-S test may be used for testing if the data are normally distributed when the sample size is larger than 50 and the distribution parameters are estimated from the data. The Lilliefors K-S test is more powerful than the chi-square test for large sample sizes and is recommended in several EPA guidance documents.

4.2.8 Recommendations

Analysts can perform tests for normality with samples as small as 3. However, the tests lack statistical power for small sample size. Therefore, for small sample sizes, it is recommended that a nonparametric statistical test (i.e., one that does not assume a distributional form of the data) be selected during Step 3 of the DQA Process in order to avoid incorrectly assuming the data are normally distributed when there is simply not enough information to test this assumption.

If the sample size is less than 50, then this guidance recommends using the Shapiro-Wilk W test, wherever practicable. The Shapiro-Wilk W test is one of most powerful tests for normality and it is recommended in several EPA guidance as the preferred test when the sample size is less than 50. This test is difficult to implement by hand but can be applied easily using the DQA DataQUEST software package (QA/G-9D, 1996). If the Shapiro-Wilk W test is not feasible, then this guidance recommends using either Filliben's statistic or the studentized range test. Filliben's statistic performs similarly to the Shapiro-Wilk test. The studentized range is a simple test to perform; however, it is not applicable for non-symmetric data with large tails. If the data are not highly skewed and the tails are not significantly large (compared to a normal distribution), the studentized range provides a simple and powerful test that can be calculated by hand. All three of these tests are included in the DataQUEST software (QA/G-9D, 1996).

If the sample size is greater than 50, this guidance recommends using either the Filliben's statistic or the studentized range test. However, if critical values for these tests (for the specific sample size) are not available, then this guidance recommends implementing either Geary's test or the Lilliefors Kolmogorov-Smirnoff test. Geary's test is easy to apply and uses standard normal tables similar to Table A-1 of Appendix A and widely available in standard textbooks. Lilliefors Kolmogorov-Smirnoff is more statistically powerful but is also more difficult to apply and uses specialized tables not readily available.

4.3 TESTS FOR TRENDS

4.3.1 Introduction

This section presents statistical tools for detecting and estimating trends in environmental data. The detection and estimation of temporal or spatial trends are important for many environmental studies or monitoring programs. In cases where temporal or spatial patterns are strong, simple procedures such as time plots or linear regression over time can reveal trends. In more complex situations, sophisticated statistical models and procedures may be needed. For example, the detection of trends may be complicated by the overlaying of long- and short-term trends, cyclical effects (e.g., seasonal or weekly systematic variations), autocorrelations, or impulses or jumps (e.g., due to interventions or procedural changes).

The graphical representations of Chapter 2 are recommended as the first step to identify possible trends. A plot of the data versus time is recommended for temporal data, as it may reveal long-term trends and may also show other major types of trends, such as cycles or impulses. A posting plot is recommended for spatial data to reveal spatial trends such as areas of high concentration or areas that were inaccessible.

For most of the statistical tools presented below, the focus is on monotonic long-term trends (i.e., a trend that is exclusively increasing or decreasing, but not both), as well as other sources of systematic variation, such as seasonality. The investigations of trend in this section are limited to one-dimensional domains, e.g., trends in a pollutant concentration over time. The current edition of this document does not address spatial trends (with 2- and 3-dimensional domains) and trends over space and time (with 3- and 4-dimensional domains), which may involve sophisticated geostatistical techniques such as kriging and require the assistance of a statistician. Section 4.3.2 discusses estimating and testing for trends using regression techniques. Section 4.3.3 discusses more robust trend estimation procedures, and section 4.3.4 discusses hypothesis tests for detecting trends under several types of situations.

4.3.2 Regression-Based Methods for Estimating and Testing for Trends

4.3.2.1 Estimating a Trend Using the Slope of the Regression Line

The classic procedures for assessing linear trends involve regression. Linear regression is a commonly used procedure in which calculations are performed on a data set containing pairs of observations (X_i, Y_i), so as to obtain the slope and intercept of a line that "best fits" the data. For temporal trends, the X_i values represent time and the Y_i values represent the observations, such as contaminant concentrations. An estimate of the magnitude of trend can be obtained by performing a regression of the data versus time (or some function of the data versus some function of time) and using the slope of the regression line as the measure of the strength of the trend.

Regression procedures are easy to apply; most scientific calculators will accept data entered as pairs and will calculate the slope and intercept of the best fitting line, as well as the correlation coefficient r (see section 2.2.4). However, regression entails several limitations and assumptions. First of all, simple linear regression (the most commonly used method) is designed to detect linear relationships between two variables; other types of regression models are generally needed to detect non-linear relationships such as cyclical or non-monotonic trends. Regression is very sensitive to extreme values (outliers), and presents difficulties in handling data below the detection limit, which are commonly encountered in environmental studies. Regression also relies on two key assumptions: normally distributed errors, and constant variance. It may be difficult or burdensome to verify these assumptions in practice, so the accuracy of the slope estimate may be

suspect. Moreover, the analyst must ensure that time plots of the data show no cyclical patterns, outlier tests show no extreme data values, and data validation reports indicate that nearly all the measurements were above detection limits. Because of these drawbacks, regression is not recommended as a general tool for estimating and detecting trends, although it may be useful as an informal, quick, and easy screening tool for identifying strong linear trends.

4.3.2.2 Testing for Trends Using Regression Methods

The limitations and assumptions associated with estimating trends based on linear regression methods apply also to other regression-based statistical tests for detecting trends. Nonetheless, for situations in which regression methods can be applied appropriately, there is a solid body of literature on hypothesis testing using the concepts of statistical linear models as a basis for inferring the existence of temporal trends. The methodology is complex and beyond the scope of this document.

For simple linear regression, the statistical test of whether the slope is significantly different from zero is equivalent to testing if the correlation coefficient is significantly different from zero. Directions for this test are given in Box 4.3-1 along with an example. This test assumes a linear relation between Y and X with independent normally distributed errors and constant variance across all X and Y values. Censored values (e.g., below the detection limit) and outliers may invalidate the tests.

Box 4.3-1: Directions for the Test for a Correlation Coefficient and an Example

Directions

STEP 1: Calculate the correlation coefficient, r (section 2.2.4).

STEP 2: Calculate the t-value $t = \dfrac{r}{\sqrt{\dfrac{1 - r^2}{n - 2}}}$.

STEP 3: Use Table A-1 of Appendix A to find the critical value $t_{1-\alpha/2}$ such that 100(1-α/2)% of the t distribution with n - 2 degrees of freedom is below $t_{1-\alpha}$. For example, if α = 0.10 and n = 17, then n-2 = 15 and $t_{1-\alpha/2}$ = 1.753. Conclude that the correlation is significantly different from zero if $|t| > t_{1-\alpha/2}$.

Example: Consider the following data set (in ppb): for Sample 1, arsenic (X) is 4.0 and lead (Y) is 8.0; for Sample 2, arsenic is 3.0 and lead is 7.0; for Sample 3, arsenic is 2.0 and lead is 7.0; and for Sample 4, arsenic is 1.0 and lead is 6.0.

STEP 1: In section 2.2.4, the correlation coefficient r for this data was calculated to be 0.949.

STEP 2: $t = \dfrac{0.949}{\sqrt{\dfrac{1 - 0.949^2}{4 - 2}}} = 4.26$

STEP 3: Using Table A-1 of Appendix A, $t_{1-\alpha/2}$ = 2.920 for a 10% level of significance and 4-2 = 2 degrees of freedom. Therefore, there appears to be a significant correlation between the two variables lead and arsenic.

4.3.3 General Trend Estimation Methods

4.3.3.1 Sen's Slope Estimator

Sen's Slope Estimate is a nonparametric alternative for estimating a slope. This approach involves computing slopes for all the pairs of ordinal time points and then using the median of these slopes as an estimate of the overall slope. As such, it is insensitive to outliers and can handle a moderate number of values below the detection limit and missing values. Assume that there are n time points (or n periods of time), and let X_i denote the data value for the ith time point. If there are no missing data, there will be n(n-1)/2 possible pairs of time points (i, j) in which i > j. The slope for such a pair is called a pairwise slope, b_{ij}, and is computed as $b_{ij} = (X_i - X_j) / (i - j)$. Sen's slope estimator is then the median of the n(n-1)/2 pairwise slopes.

If there is no underlying trend, then a given X_i is as likely to be above another X_j as it is below. Hence, if there is no underlying trend, there would be an approximately equal number of positive and negative slopes, and thus the median would be near zero. Due to the number of calculations required, Sen's estimator is rarely calculated by hand and directions are not given in this document. However, the estimator is contained in the DQA DataQUEST software package (QA/G-9D, 1996).

4.3.3.2 Seasonal Kendall Slope Estimator

If the data exhibit cyclic trends, then Sen's slope estimator can be modified to account for the cycles. For example, if data are available for each month for a number of years, 12 separate sets of slopes would be determined (one for each month of the year); similarly, if daily observations exhibit weekly cycles, seven sets of slopes would be determined, one for each day of the week. In these estimates, the above pairwise slope is calculated for each time period and the median of all of the slopes is an estimator of the slope for a long-term trend. This is known as the seasonal Kendall slope estimator. Because of the number of calculations required, this estimator is rarely calculated by hand so directions are not given in this document. The seasonal Kendall slope estimator is contained in the DataQUEST software package (QA/G-9D, 1996).

4.3.4 Hypothesis Tests for Detecting Trends

Most of the trend tests treated in this section involve the Mann-Kendall test or extensions of it. The Mann-Kendall test does not assume any particular distributional form and accommodates trace values or values below the detection limit by assigning them a common value. The test can also be modified to deal with multiple observations per time period and generalized to deal with multiple sampling locations and seasonality.

4.3.4.1 One Observation per Time Period for One Sampling Location

The Mann-Kendall test involves computing a statistic S, which is the difference between the number of pairwise slopes (as described in 4.3.3.1) that are positive minus the number that are negative. If S is a large positive value, then there is evidence of an increasing trend in the data. If S is a large negative value, then there is evidence of a decreasing trend in the data. The null hypothesis or baseline condition for this test is that there is no temporal trend in the data values, i.e., "H_0: no trend". The alternative condition or hypothesis will usually be either "H_A: upward trend" or "H_A: downward trend."

The basic Mann-Kendall trend test involves listing the observations in temporal order, and computing all differences that may be formed between measurements and earlier measurements, as depicted

in Box 4.3-2. The test statistic is the difference between the number of strictly positive differences and the number of strictly negative differences. If there is an underlying upward trend, then these differences will tend to be positive and a sufficiently large value of the test statistic will suggest the presence of an upward trend. Differences of zero are not included in the test statistic (and should be avoided, if possible, by recording data to sufficient accuracy). The steps for conducting the Mann-Kendall test for small sample sizes (i.e., less than 10) are contained in Box 4.3-3 and an example is contained in Box 4.3-4.

For sample sizes greater than 10, a normal approximation to the Mann-Kendall test is quite accurate. Directions for this approximation are contained in Box 4.3-5 and an example is given in Box 4.3-6. Tied observations (i.e., when two or more measurements are equal) degrade the statistical power and should be avoided, if possible, by recording the data to sufficient accuracy.

Box 4.3-2: "Upper Triangular" Data for Basic Mann-Kendall Trend Test with a Single Measurement at Each Time Point

Data Table

Original Time Measurements	t_1 X_1	t_2 X_2	t_3 X_3	t_4 X_4	. . . . .	t_{n-1} X_{n-1}	t_n X_n	(time from earliest to latest) (actual values recorded)
X_1		X_2-X_1	X_3-X_1	X_4-X_1	. . .	$X_{n-1}-X_1$	X_n-X_1	
X_2			X_3-X_2	X_4-X_2	. . .	$X_{n-1}-X_2$	X_n-X_2	
. . .						. . .	. . .	
X_{n-2}						$X_{n-1}-X_{n-2}$	X_n-X_{n-2}	
X_{n-1}							X_n-X_{n-1}	

After performing the subtractions this table converts to:

Original Time Measurements	t_1 X_1	t_2 X_2	t_3 X_3	t_4 X_4	. . . . .	t_{n-1} X_{n-1}	t_n X_n	# of + Differences (>0)	# of - Differences (<0)
X_1		Y_{21}	Y_{31}	Y_{41}	. . .	$Y_{(n-1)1}$	Y_{n1}		
X_2			Y_{32}	Y_{42}	. . .	$Y_{(n-1)2}$	Y_{n2}		
. . .						. . .	. . .		
X_{n-2}						$Y_{(n-1)(n-2)}$	$Y_{n(n-2)}$		
X_{n-1}							$Y_{n\,(n-1)}$		
NOTE: $X_{ij}-Y_k=0$ do not contribute to either total and are discarded.								Total # >0	Total # <0

where Y_{ik} = sign (X_i-X_k) = + if $X_i - X_k > 0$
= 0 if $X_i - X_k = 0$
= - if $X_i - X_k < 0$

Box 4.3-3: Directions for the Mann-Kendall Trend Test for Small Sample Sizes

If the sample size is less than 10 and there is only one datum per time period, the Mann-Kendall Trend Test for small sample sizes may be used.

STEP 1: List the data in the order collected over time: $X_1, X_2, ..., X_n$, where X_i is the datum at time t_i. Assign a value of DL/2 to values reported as below the detection limit (DL). Construct a "Data Matrix" similar to the top half of Box 4.3-2.

STEP 2: Compute the sign of all possible differences as shown in the bottom portion of Box 4.3-2.

STEP 3: Compute the Mann-Kendall statistic S, which is the number of positive signs minus the number of negative signs in the triangular table: S = (number of + signs) - (number of - signs).

STEP 4: Use Table A-11 of Appendix A to determine the probability p using the sample size n and the absolute value of the statistic S. For example, if n=5 and S=8, p=0.042.

STEP 5: For testing the null hypothesis of no trend against H_1 (upward trend), reject H_0 if $S > 0$ and if $p < \alpha$. For testing the null hypothesis of no trend against H_2 (downward trend), reject H_0 if $S < 0$ and if $p < \alpha$.

Box 4.3-4: An Example of Mann-Kendall Trend Test for Small Sample Sizes

Consider 5 measurements ordered by the time of their collection: 5, 6, 11, 8, and 10. This data will be used to test the null hypothesis, H_0: no trend, versus the alternative hypothesis H_1 of an upward trend at an $\alpha = 0.05$ significance level.

STEP 1: The data listed in order by time are: 5, 6, 11, 8, 10.

STEP 2: A triangular table (see Box 4.3-2) was used to construct the possible differences. The sum of signs of the differences across the rows are shown in the columns 7 and 8.

Time	1	2	3	4	5	No. of +	No. of
Data	5	6	11	8	10	Signs	- Signs
5		+	+	+	+	4	0
6			+	+	+	3	0
11				-	-	0	2
8					+	1	0
						8	2

STEP 3: Using the table above, S = 8 - 2 = 6.

STEP 4: From Table A-11 of Appendix A for n = 5 and S = 6, p = 0.117.

STEP 5: Since $S > 0$ but $p = 0.117 \not< 0.05$, the null hypothesis is not rejected. Therefore, there is not enough evidence to conclude that there is an increasing trend in the data.

Box 4.3-5: Directions for the Mann-Kendall Procedure Using Normal Approximation

If the sample size is 10 or more, a normal approximation to the Mann-Kendall procedure may be used.

STEP 1: Complete steps 1, 2, and 3 of Box 4.3-3.

STEP 2: Calculate the variance of S: $V(S) = \dfrac{n(n-1)(2n+5)}{18}$.

If ties occur, let g represent the number of tied groups and w_p represent the number of data points in the pth group. The variance of S is: $V(S) = \dfrac{1}{18}\left[n(n-1)(2n+5) - \sum_{p=1}^{g} w_p(w_p-1)(2w_p+5)\right]$

STEP 4: Calculate $Z = \dfrac{S-1}{[V(S)]^{1/2}}$ if S > 0, Z = 0 if S = 0, or $Z = \dfrac{S+1}{[V(S)]^{1/2}}$ if S < 0.

STEP 5: Use Table A-1 of Appendix A to find the critical value $z_{1-\alpha}$ such that 100(1-α)% of the normal distribution is below $z_{1-\alpha}$. For example, if α=0.05 then $z_{1-\alpha}$=1.645.

STEP 6: For testing the hypothesis, H_0 (no trend) against 1) H_1 (an upward trend) – reject H_0 if Z is greater than $z_{1-\alpha}$, or 2) H_2 (a downward trend) – reject H_0 if Z < 0 and the absolute value of Z is greater than $z_{1-\alpha}$.

Box 4.3-6: An Example of Mann-Kendall Trend Test by Normal Approximation

A test for an upward trend with α=.05 will be based on the 11 weekly measurements shown below.

STEP 1: Using Box 4.3-2, a triangular table was used to construct the possible differences. A zero has been used if the difference is zero, a "+" sign if the difference is positive, and a "-" sign if the difference is negative.

Week	1	2	3	4	5	6	7	8	9	10	11	No. of	No. of
Data	10	10	10	5	10	20	18	17	15	24	15	+ Signs	- Signs
10		0	0	-	0	+	+	+	+	+	+	6	1
10			0	-	0	+	+	+	+	+	+	6	1
10				-	0	+	+	+	+	+	+	6	1
5					+	+	+	+	+	+	+	7	0
10						+	+	+	+	+	+	6	0
20							-	-	-	+	-	1	4
18								-	-	+	-	1	3
17									-	+	-	1	2
15										+	0	1	0
24											-	0	1
												35	13

STEP 2: S = (sum of + signs) - (sum of - signs) = 35 - 13 = 22

STEP 3: There are several observations tied at 10 and 15. Thus, the formula for tied values will be used. In this formula, g=2, t_1=4 for tied values of 10, and t_2=2 for tied values of 15.

$$V(S) = \frac{1}{18}\left[11(11-1)(2(11)+5) - [4(4-1)(2(4)+5) + 2(2-1)(2(2)+5)]\right] = 155.33$$

STEP 4: Since S is positive: $Z = \dfrac{S-1}{[V(S)]^{1/2}} = \dfrac{22-1}{(155.33)^{1/2}} = \dfrac{20}{12.46} = 1.605$

STEP 5: From Table A-1 of Appendix A, $z_{1-.05}$=1.645.

STEP 6: H_1 is the alternative of interest. Therefore, since 1.605 is not greater than 1.645, H_0 is not rejected. Therefore, there is not enough evidence to determine that there is an upward trend.

4.3.4.2 Multiple Observations per Time Period for One Sampling Location

Often, more than one sample is collected for each time period. There are two ways to deal with multiple observations per time period. One method is to compute a summary statistic, such as the median, for each time period and to apply one of the Mann-Kendall trend tests of section 4.3.4.1 to the summary statistic. Therefore, instead of using the individual data points in the triangular table, the summary statistic would be used. Then the steps given in Box 4.3-3 and 4.3-5 could be applied to the summary statistics.

An alternative approach is to consider all the multiple observations within a given time period as being essentially equal (i.e., tied) values within that period. The S statistic is computed as before with n being the total of all observations. The variance of the S statistic (previously calculated in step 2) is changed to:

$$VAR(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p=1}^{g} w_p(w_p-1)(2w_p+5) - \sum_{q=1}^{h} u_q(u_q-1)(2u_q+5)\right]$$

$$+ \frac{\sum_{p=1}^{g} w_p(w_p-1)(w_p-2)\sum_{q=1}^{h} u_q(u_q-1)(u_q-2)}{9n(n-1)(n-2)} + \frac{\sum_{p=1}^{g} w_p(w_p-1)\sum_{q=1}^{h} u_q(u_q-1)}{2n(n-1)}$$

where g represents the number of tied groups, w_p represents the number of data points in the pth group, h is the number of time periods which contain multiple data, and u_q is the sample size in the qth time period.

The preceding variance formula assumes that the data are not correlated. If correlation within single time periods is suspected, it is preferable to use a summary statistic (e.g., the median) for each time period and to then apply either Box 4.3-3 or Box 4.3-5 to the summary statistics.

4.3.4.3 Multiple Sampling Locations with Multiple Observations

The preceding methods involve a single sampling location (station). However, environmental data often consist of sets of data collected at several sampling locations (see Box 4.3-7). For example, data are often systematically collected at several fixed sites on a lake or river, or within a region or basin. The data collection plan (or experimental design) must be systematic in the sense that approximately the same sampling times should be used at all locations. In this situation, it is desirable to express the results by an overall regional summary statement across all sampling locations. However, there must be consistency in behavioral characteristics across sites over time in order for a single summary statement to be valid across all sampling locations. A useful plot to assess the consistency requirement is a single time plot (section 2.3.8.1) of the measurements from all stations where a different symbol is used to represent each station.

If the stations exhibit approximately steady trends in the same direction (upward or downward), with comparable slopes, then a single summary statement across stations is valid and this implies two relevant sets of hypotheses should be investigated:

Comparability of stations. H_0: Similar dynamics affect all K stations vs. H_A: At least two stations exhibit different dynamics.

Testing for overall monotonic trend. H_0*: Contaminant levels do not change over time vs. H_A': There is an increasing (or decreasing) trend consistently exhibited across all stations.

Therefore, the analyst must first test for homogeneity of stations, and then, if homogeneity is confirmed, test for an overall monotonic trend.

Ideally, the stations in Box 4.3-7 should have equal numbers. However, the numbers of observations at the stations can differ slightly, because of isolated missing values, but the overall time periods spanned must be similar. This guidance recommends that for less than 3 time periods, an equal number of observations (a balanced design) is required. For 4 or more time periods, up to 1 missing value per sampling location may be tolerated.

a. One Observation per Time Period. When only one measurement is taken for each time period for each station, a generalization of the Mann-Kendall statistic can be used to test the above hypotheses. This procedure is described in Box 4.3-8.

b. Multiple Observations per Time Period. If multiple measurements are taken at some times and station, then the previous approaches are still applicable. However, the variance of the statistic S_k must be calculated using the equation for calculating V(S) given in section 4.3.4.2. Note that S_k is computed for each station, so n, w_p, g, h, and u_q are all station-specific.

Box 4.3-7: Data for Multiple Times and Multiple Stations

Let i = 1, 2, ..., n represent time, k = 1, 2, ..., K represent sampling locations, and X_{ik} represent the measurement at time i for location k. This data can be summarized in matrix form, as shown below.

		Stations			
		1	2	. . .	K
	1	X_{11}	X_{12}	. . .	X_{1K}
	2	X_{21}	X_{22}	. . .	X_{2K}
Time	.	.	.	. . .	.
	.	.	.	. . .	.
	.	.	.	. . .	.
	n	X_{n1}	X_{n2}	. . .	X_{nK}
		S_1	S_2	. . .	S_K
		$V(S_1)$	$V(S_2)$	. . .	$V(S_K)$
		Z_1	Z_2	. . .	Z_K

where S_k = Mann-Kendall statistic for station k (see STEP 3, Box 4.3-3),
$V(S_k)$ = variance for S statistic for station k (see STEP 2, Box 4.3-5), and
$Z_k = S_k / \sqrt{VAR(S_k)}$

Box 4.3-8: Testing for Comparability of Stations and an Overall Monotonic Trend

Let i = 1, 2, ..., n represent time, k = 1, 2, ..., K represent sampling locations, and X_{ik} represent the measurement at time i for location k. Let α represent the significance level for testing homogeneity and α^* represent the significance level for testing for an overall trend.

STEP 1: Calculate the Mann-Kendall statistic S_k and its variance $V(S_k)$ for each of the K stations using the methods of section 4.3.4.1, Box 4.3-5.

STEP 2: For each of the K stations, calculate $Z_k = S_k / \sqrt{V(S_k)}$.

STEP 3: Calculate the average $\bar{Z} = \sum_{k=1}^{K} Z_k / K$.

STEP 4: Calculate the homogeneity chi-square statistic $\chi_h^2 = \sum_{k=1}^{K} Z_k^2 - K\,\bar{Z}^2$.

STEP 5: Using a chi-squared table (Table A-8 of Appendix A), find the critical value for χ^2 with (K-1) degrees of freedom at an α significance level. For example, for a significance level of 5% and 5 degrees of freedom, $\chi^2_{(5)} = 11.07$, i.e., 11.07 is the cut point which puts 5% of the probability in the upper tail of a chi-square variable with 5 degrees of freedom.

STEP 6: If $\chi_h^2 \le \chi^2_{(K-1)}$, there are comparable dynamics across stations at significance level α. Go to Step 7.

If $\chi_h^2 > \chi^2_{(K-1)}$, the stations are not homogeneous (i.e., different dynamics at different stations) at the significance level α. Therefore, individual α^*-level Mann-Kendall tests should be conducted at each station using the methods presented in section 4.3.4.1.

STEP 7: Using a chi-squared table (Table A-8 of Appendix A), find the critical value for χ^2 with 1 degree of freedom at an α significance level. If

$$K\,\bar{Z}^2 > \chi^2_{(1)},$$

then reject H_0^* and conclude that there is a significant (upward or downward) monotonic trend across all stations at significance level α^*. The signs of the S_k indicate whether increasing or decreasing trends are present. If

$$K\,\bar{Z}^2 \le \chi^2_{(1)},$$

there is not significant evidence at the α' level of a monotonic trend across all stations. That is, the stations appear approximately stable over time.

4.3.4.4 One Observation for One Station with Multiple Seasons

Temporal data are often collected over extended periods of time. Within the time variable, data may exhibit periodic cycles, which are patterns in the data that repeat over time (e.g., the data may rise and fall regularly over the months in a year or the hours in a day). For example, temperature and humidity may change with the season or month, and may affect environmental measurements. (For more information on seasonal cycles, see section 2.3.8). In the following discussion, the term season represents one time point in the periodic cycle, such as a month within a year or an hour within a day.

If seasonal cycles are anticipated, then two approaches for testing for trends are the seasonal Kendall test and Sen's test for trends. The seasonal Kendall test may be used for large sample sizes, and Sen's test for trends may be used for small sample sizes. If different seasons manifest similar slopes (rates of change) but possibly different intercepts, then the Mann-Kendall technique of section 4.3.4.3 is applicable, replacing time by year and replacing station by season.

The seasonal Kendall test, which is an extension of the Mann-Kendall test, involves calculating the Mann-Kendall test statistic, S, and its variance separately for each “season” (e.g., month of the year, day of the week). The sum of the S's and the sum of their variances are then used to form an overall test statistic that is assumed to be approximately normally distributed for larger size samples.

For data at a single site, collected at multiple seasons within multiple years, the techniques of section 4.3.4.3 can be applied to test for homogeneity of time trends across seasons. The methodology follows Boxes 4.3-7 and 4.3-8 exactly except that “station” is replaced by “season” and the inferences refer to seasons.

4.3.5 A Discussion on Tests for Trends

This section discusses some further considerations for choosing among the many tests for trends. All of the nonparametric trend tests and estimates use ordinal time (ranks) rather than cardinal time (actual time values, such as month, day or hour) and this restricts the interpretation of measured trends. All of the Mann-Kendall (MK) Trend Tests presented are based on certain pairwise differences in measurements at different time points. The only information about these differences that is used in the MK calculations is their signs (i.e., whether they are positive or negative) and therefore are generalizations of the sign test. MK calculations are relatively easy and simply involve counting the number of cases in which X_{i+j} exceeds X_i and the number of cases in which X_i exceeds X_{i+j}. Information about magnitudes of these differences is not used by MK methods and this can adversely affect the statistical power when only limited amounts of data are available.

There are, however, nonparametric methods based on ranks that takes such magnitudes into account and still retains the benefit of robustness to outliers. These procedures can be thought of as replacing the data by their ranks and then conducting parametric analyses. These include the Wilcoxon rank sum test and its many generalizations. These methods are more resistant to outliers than parametric methods; a point can be no more extreme than the smallest or largest value.

Rank-based methods, which make fuller use of the information in the data than MK methods, are not as robust with respect to outliers as the sign and MK tests. They are, however, more statistically powerful than the sign test and MK methods; the Wilcoxon test being a case in point. If the data are random samples from normal distributions with equal variances, then the sign test requires approximately 1.225 times as many observations as the Wilcoxon rank sum test to achieve a given power at a given significance level. This kind of tradeoff between power and robustness exemplifies the analyst's evaluation process leading to the selection of the best statistical procedure for the current situation. Further statistical tests will be developed in future editions of this guidance.

4.4 OUTLIERS

4.4.1 Background

Outliers are measurements that are extremely large or small relative to the rest of the data and, therefore, are suspected of misrepresenting the population from which they were collected. Outliers may result from transcription errors, data-coding errors, or measurement system problems such as instrument breakdown. However, outliers may also represent true extreme values of a distribution (for instance, hot spots) and indicate more variability in the population than was expected. Not removing true outliers and removing false outliers both lead to a distortion of estimates of population parameters.

Statistical outlier tests give the analyst probabilistic evidence that an extreme value (potential outlier) does not "fit" with the distribution of the remainder of the data and is therefore a statistical outlier. These tests should only be used to *identify* data points that require further investigation. The tests alone cannot determine whether a statistical outlier should be discarded or corrected within a data set; this decision should be based on judgmental or scientific grounds..

There are 5 steps involved in treating extreme values or outliers:

1. Identify extreme values that may be potential outliers;
2. Apply statistical test;
3. Scientifically review statistical outliers and decide on their disposition;
4. Conduct data analyses with and without statistical outliers; and
5. Document the entire process.

Potential outliers may be identified through the graphical representations of Chapter 2 (step 1 above). Graphs such as the box and whisker plot, ranked data plot, normal probability plot, and time plot can all be used to identify observations that are much larger or smaller than the rest of the data. If potential outliers are identified, the next step is to apply one of the statistical tests described in the following sections. Section 4.4.2 provides recommendations on selecting a statistical test for outliers.

If a data point is found to be an outlier, the analyst may either: 1) correct the data point; 2) discard the data point from analysis; or 3) use the data point in all analyses. This decision should be based on scientific reasoning *in addition to* the results of the statistical test. For instance, data points containing transcription errors should be corrected, whereas data points collected while an instrument was malfunctioning may be discarded. One should never discard an outlier based solely on a statistical test. Instead, the decision to discard an outlier should be based on some scientific or quality assurance basis. Discarding an outlier from a data set should be done with extreme caution, particularly for environmental data sets, which often contain legitimate extreme values. If an outlier is discarded from the data set, all statistical analysis of the data should be applied to both the full and truncated data set so that the effect of discarding observations may be assessed. If scientific reasoning does not explain the outlier, it should not be discarded from the data set.

If any data points are found to be statistical outliers through the use of a statistical test, this information will need to be documented along with the analysis of the data set, regardless of whether any data points are discarded. If no data points are discarded, document the identification of any "statistical" outliers by documenting the statistical test performed and the possible scientific reasons investigated. If any data points are discarded, document each data point, the statistical test performed, the scientific reason for

discarding each data point, and the effect on the analysis of deleting the data points. This information is critical for effective peer review.

4.4.2 Selection of a Statistical Test

There are several statistical tests for determining whether or not one or more observations are statistical outliers. Step by step directions for implementing some of these tests are described in sections 4.4.3 through 4.4.6. Section 4.4.7 describes statistical tests for multivariate outliers.

Sample Size	Test	Section	Assumes Normality	Multiple Outliers	Data-QUEST
$n \leq 25$	Extreme Value Test	4.4.3	Yes	No/Yes	Yes
$n \leq 50$	Discordance Test	4.4.4	Yes	No	Yes
$n \geq 25$	Rosner's Test	4.4.5	Yes	Yes	Yes
$n \geq 50$	Walsh's Test	4.4.6	No	Yes	Yes

Table 4.4-1. Recommendations for Selecting a Statistical Test for Outliers

If the data are normally distributed, this guidance recommends applying Rosner's test (Box 4.4-5) when the sample size is greater than 25 and the Extreme Value test (Box 4.4-1) when the sample size is less than 25. If only one outlier is suspected, then the Discordance test (Box 4.4-3) may be substituted for either of these tests. If the data are not normally distributed, or if the data cannot be transformed so that the transformed data are normally distributed, then the analyst should either apply a nonparametric test (such as Walsh's test in Box 4.4-7) or consult a statistician.

4.4.3 Extreme Value Test (Dixon's Test)

Dixon's Extreme Value test can be used to test for statistical outliers when the sample size is less than or equal to 25. This test considers both extreme values that are much smaller than the rest of the data (case 1) and extreme values that are much larger than the rest of the data (case 2). This test assumes that the data without the suspected outlier are normally distributed; therefore, it is necessary to perform a test for normality on the data without the suspected outlier before applying this test. If the data are not normally distributed, either transform the data, apply a different test, or consult a statistician. Directions for the Extreme Value test are contained in Box 4.4-1; an example of this test is contained in Box 4.4-2. The Extreme Value test is contained in the DQA DataQUEST software package (QA/G-9D, 1996).

This guidance recommends using this test when only one outlier is suspected in the data. If more than one outlier is suspected, the Extreme Value test may lead to masking where two or more outliers close in value “hide” one another. Therefore, if the analyst decides to use the Extreme Value test for multiple outliers, apply the test to the least extreme value first.

Box 4.4-1: Directions for the Extreme Value Test (Dixon's Test)

STEP 1: Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ represent the data ordered from smallest to largest. Check that the data without the suspect outlier are normally distributed, using one of the methods of section 4.2. If normality fails, either transform the data or apply a different outlier test.

STEP 2: $X_{(1)}$ is a Potential Outlier (case 1): Compute the test statistic C, where

$$C = \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}} \text{ for } 3 \le n \le 7, \quad C = \frac{X_{(3)} - X_{(1)}}{X_{(n-1)} - X_{(1)}} \text{ for } 11 \le n \le 13,$$

$$C = \frac{X_{(2)} - X_{(1)}}{X_{(n-1)} - X_{(1)}} \text{ for } 8 \le n \le 10, \quad C = \frac{X_{(3)} - X_{(1)}}{X_{(n-2)} - X_{(1)}} \text{ for } 14 \le n \le 25.$$

STEP 3: If C exceeds the critical value from Table A-3 of Appendix A for the specified significance level α, $X_{(1)}$ is an outlier and should be further investigated.

STEP 4: $X_{(n)}$ is a Potential Outlier (case 2): Compute the test statistic C, where

$$C = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}} \text{ for } 3 \le n \le 7, \quad C = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(2)}} \text{ for } 11 \le n \le 13,$$

$$C = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(2)}} \text{ for } 8 \le n \le 10, \quad C = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(3)}} \text{ for } 14 \le n \le 25$$

STEP 5: If C exceeds the critical value from Table A-3 of Appendix A for the specified significance level α, $X_{(n)}$ is an outlier and should be further investigated.

Box 4.4-2: An Example of the Extreme Value Test (Dixon's Test)

The data in order of magnitude from smallest to largest are: 82.39, 86.62, 91.72, 98.37, 103.46, 104.93, 105.52, 108.21, 113.23, and 150.55 ppm. Because the largest value (150.55) is much larger than the other values, it is suspected that this data point might be an outlier.

STEP 1: A normal probability plot of the data shows that there is no reason to suspect that the data (without the extreme value) are not normally distributed. The studentized range test (section 4.2.6) also shows that there is no reason to suspect that the data are not normally distributed. Therefore, the Extreme Value test may be used to determine if the largest data value is an outlier.

STEP 4: $$C = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(2)}} = \frac{150.55 - 113.23}{150.55 - 86.62} = \frac{37.32}{63.93} = 0.584$$

STEP 5: Since C = 0.584 > 0.477 (from Table A-3 of Appendix A with n=10), there is evidence that $X_{(n)}$ is an outlier at a 5% significance level and should be further investigated.

4.4.4 Discordance Test

The Discordance test can be used to test if one extreme value is an outlier. This test considers two cases: 1) where the extreme value (potential outlier) is the smallest value of the data set, and 2) where the extreme value (potential outlier) is the largest value of the data set. The Discordance test assumes that the data are normally distributed; therefore, it is necessary to perform a test for normality before applying this test. If the data are not normally distributed either transform the data, apply a different test, or consult a statistician. Note that the test assumes that the data without the outlier are normally distributed; therefore, the test for normality should be performed without the suspected outlier. Directions and an example of the Discordance test are contained in Box 4.4-3 and 4.4-4, respectively. Table A-4 of Appendix A contains critical values for this test for $n \leq 50$.

Box 4.4-3: Directions for the Discordance Test

STEP 1: Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ represent the data ordered from smallest to largest. Check that the data without the suspect outlier are normally distributed, using one of the methods of section 4.2. If normality fails, either transform the data or apply a different outlier test.

STEP 2: Compute the sample mean, $\bar{x}$ (section 2.2.2), and the sample standard deviation, s (section 2.2.3). If the minimum value $X_{(1)}$ is a suspected outlier, perform Steps 3 and 4. If the maximum value $X_{(n)}$ is a suspected outlier, perform Steps 5 and 6.

STEP 3: <u>If $X_{(1)}$ is a Potential Outlier (case 1)</u>: Compute the test statistic $D = \dfrac{\bar{X} - X_{(1)}}{s}$

STEP 4: If D exceeds the critical value from Table A-4, $X_{(1)}$ is an outlier and should be further investigated.

STEP 5: <u>If $X_{(n)}$ is a Potential Outlier (case 2)</u>: Compute the test statistic $D = \dfrac{X_{(n)} - \bar{X}}{s}$

STEP 6: If D exceeds the critical value from Table A-4, $X_{(n)}$ is an outlier and should be further investigated.

Box 4.4-4: An Example of the Discordance Test

The ordered data are 82.39, 86.62, 91.72, 98.37, 103.46, 104.93, 105.52, 108.21, 113.23, and 150.55 ppm. Because the largest value of this data set (150.55) is much larger than the rest, it may be an outlier.

STEP 1: A normal probability plot of the data shows that there is no reason to suspect that the data (without the extreme value) are not normally distributed. The studentized range test (section 4.2.6) also shows that there is no reason to suspect that the data are not normally distributed. Therefore, the Discordance test may be used to determine if the largest data value is an outlier.

STEP 2: $\bar{x}$ = 104.5 ppm and s = 18.922 ppm.

STEP 5: $D = \dfrac{X_{(n)} - \bar{X}}{s} = \dfrac{150.55 - 104.50}{18.92} = 2.43$

STEP 6: Since D = 2.43 > 2.176 (from Table A-4 of Appendix A with n = 10), there is evidence that $X_{(n)}$ is an outlier at a 5% significance level and should be further investigated.

4.4.5 Rosner's Test

A parametric test developed by Rosner can be used to detect up to 10 outliers for sample sizes of 25 or more. This test assumes that the data are normally distributed; therefore, it is necessary to perform a test for normality before applying this test. If the data are not normally distributed either transform the data, apply a different test, or consult a statistician. Note that the test assumes that the data without the outlier are normally distributed; therefore, the test for normality may be performed without the suspected outlier. Directions for Rosner's test are contained in Box 4.4-5 and an example is contained in Box 4.4-6. This test is also contained in the DQA DataQUEST software package (QA/G-9D, 1996).

Rosner's test is not as easy to apply as the preceding tests. To apply Rosner's test, first determine an upper limit r_0 on the number of outliers ($r_0 \leq 10$), then order the r_0 extreme values from most extreme to least extreme. Rosner's test statistic is then based on the sample mean and sample standard deviation computed without the $r = r_0$ extreme values. If this test statistic is greater than the critical value given in Table A-5 of Appendix A, there are r_0 outliers. Otherwise, the test is performed again without the $r = r_0 - 1$ extreme values. This process is repeated until either Rosner's test statistic is greater than the critical value or r = 0.

Box 4.4-5: Directions for Rosner's Test for Outliers

STEP 1: Let $X_1, X_2, \ldots, X_n$ represent the ordered data points. By inspection, identify the maximum number of possible outliers, r_0. Check that the data are normally distributed, using one of the methods of section 4.2.

STEP 2: Compute the sample mean $\bar{x}$, and the sample standard deviation, s, for <u>all</u> the data. Label these values $\bar{x}^{(0)}$ and $s^{(0)}$, respectively. Determine the observation farthest from $\bar{x}^{(0)}$ and label this observation $y^{(0)}$. Delete $y^{(0)}$ from the data and compute the sample mean, labeled $\bar{x}^{(1)}$, and the sample standard deviation, labeled $s^{(1)}$. Then determine the observation farthest from $\bar{x}^{(1)}$ and label this observation $y^{(1)}$. Delete $y^{(1)}$ and compute $\bar{x}^{(2)}$ and $s^{(2)}$. Continue this process until r_0 extreme values have been eliminated.

In summary, after the above process the analyst should have

$$[\bar{X}^{(0)}, s^{(0)}, y^{(0)}]; [\bar{X}^{(1)}, s^{(1)}, y^{(1)}]; \ldots, [\bar{X}^{(r_0-1)}, s^{(r_0-1)}, y^{(r_0-1)}] \quad \text{where}$$

$$\bar{X}^{(i)} = \frac{1}{n-i}\sum_{j=1}^{n-i} x_j, \quad s^{(i)} = [\frac{1}{n-i}\sum_{j=1}^{n-i}(x_j - \bar{x}^{(i)})^2]^{1/2},$$ and $y^{(i)}$ is the farthest value

from $\bar{x}^{(i)}$. (Note, the above formulas for $\bar{x}^{(i)}$ and $s^{(i)}$ assume that the data have been renumbered after each observation is deleted.)

STEP 3: To test if there are 'r' outliers in the data, compute: $R_r = \dfrac{|y^{(r-1)} - \bar{x}^{(r-1)}|}{s^{(r-1)}}$ and compare

R_r to λ_r in Table A-5 of Appendix A. If $R_r \geq \lambda_r$, conclude that there are r outliers.

First, test if there are r_0 outliers (compare $R_{r_{o-1}}$ to $\lambda_{r_{o-1}}$). If not, test if there are $r_0 - 1$ outliers (compare $R_{r_{o-1}}$ to $\lambda_{r_{o-1}}$). If not, test if there are $r_0 - 2$ outliers, and continue, until either it is determined that there are a certain number of outliers or that there are no outliers at all.

Box 4.4-6: An Example of Rosner's Test for Outliers

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 shows that there is no reason to suspect that the data (without the suspect outliers) are not normally distributed. In addition, this graph identified four potential outliers: 2.07, 40.55, 275.36, and 395.67. Therefore, Rosner's test will be applied to see if there are 4 or fewer ($r_0 = 4$) outliers.

STEP 2: First the sample mean and sample standard deviation were computed for the entire data set ($\bar{x}^{(0)}$ and $s^{(0)}$). Using subtraction, it was found that 395.67 was the farthest data point from $\bar{x}^{(0)}$, so $y^{(0)} = 395.67$. Then 395.67 was deleted from the data and the sample mean, $\bar{x}^{(1)}$, and the sample standard deviation, $s^{(1)}$, were computed. Using subtraction, it was found that 2.07 was the farthest value from $\bar{x}^{(1)}$. This value was then dropped from the data and the process was repeated again on 40.55 to yield $\bar{x}^{(2)}$, $s^{(2)}$, and $y^{(2)}$ and $\bar{x}^{(3)}$, $s^{(3)}$, and $y^{(3)}$. These values are summarized below.

i	$\bar{x}^{(i)}$	$s^{(i)}$	$y^{(i)}$
0	169.923	75.133	395.67
1	162.640	63.872	2.07
2	167.993	57.460	40.55
3	172.387	53.099	275.36

STEP 3: To apply Rosner's test, it is first necessary to test if there are 4 outliers by computing

$$R_4 = \frac{|y^{(3)} - \bar{x}^{(3)}|}{s^{(3)}} = \frac{|275.36 - 172.387|}{53.099} = 1.939$$

and comparing R_4 to λ_4 in Table A-5 of Appendix A with n = 32. Since $R_4 = 1.939 \ngeq \lambda_4 = 2.89$, there are not 4 outliers in the data set. Therefore, it will next be tested if there are 3 outliers by computing

$$R_3 = \frac{|y^{(2)} - \bar{x}^{(2)}|}{s^{(2)}} = \frac{|40.55 - 167.993|}{57.460} = 2.218$$

and comparing R_3 to λ_3 in Table A-5 with n = 32. Since $R_3 = 2.218 \ngeq \lambda_3 = 2.91$, there are not 3 outliers in the data set. Therefore, it will next be tested if there are 2 outliers by computing

$$R_2 = \frac{|y^{(1)} - \bar{x}^{(1)}|}{s^{(1)}} = \frac{|2.07 - 162.640|}{63.872} = 2.514$$

and comparing R_2 to λ_2 in Table A-5 with n = 32. Since $R_2 = 2.514 \ngeq \lambda_3 = 2.92$, there are not 2 outliers in the data set. Therefore, it will next be tested if there is 1 outlier by computing

$$R_1 = \frac{|y^{(0)} - \bar{x}^{(0)}|}{s^{(0)}} = \frac{|395.67 - 169.923|}{75.133} = 3.005$$

and comparing R_1 to λ_1 in Table A-5 with n = 32. Since $R_1 = 3.005 > \lambda_1 = 2.94$, there is evidence at a 5% significance level that there is 1outlier in the data set. Therefore, observation 355.67 is a statistical outlier and should be further investigated.

4.4.6 Walsh's Test

A nonparametric test was developed by Walsh to detect multiple outliers in a data set. This test requires a large sample size: n > 220 for a significance level of $\alpha = 0.05$, and n > 60 for a significance level of $\alpha = 0.10$. However, since the test is a nonparametric test, it may be used whenever the data are not normally distributed. Directions for the test by Walsh for large sample sizes are given in Box 4.4-7. This test is also contained in the DQA DataQUEST software package (QA/G-9D, 1996).

Box 4.4-7: Directions for Walsh's Test for Large Sample Sizes

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ represent the data ordered from smallest to largest. If $n \leq 60$, do not apply this test. If $60 < n \leq 220$, then $\alpha = 0.10$. If $n > 220$, then $\alpha = 0.05$.

STEP 1: Identify the number of possible outliers, r. Note that r can equal 1.

STEP 2: Compute $c = \sqrt{2n},\ k = r + c,\ b^2 = 1/\alpha,\ and\ \ a = \dfrac{1 + b\sqrt{(c - b^2)/(c-1)}}{c - b^2 - 1}.$

Round k to the smallest integer less than k.

STEP 3: The r smallest points are outliers (with a α% level of significance) if

$$x_{(r)} - (1+a)x_{(r-1)} + a x_{(k)} < 0$$

STEP 4: The r largest points are outliers (with a α% level of significance) if

$$x_{(n+1-r)} - (1+a)x_{(n-r)} + a x_{(n+1-k)} > 0$$

STEP 5: If both of the inequalities are true, then both small and large outliers are indicated.

4.4.7 Multivariate Outliers

Multivariate analysis, such as factor analysis and principal components analysis, involves the analysis of several variables simultaneously. Outliers in multivariate analysis are then values that are extreme in relationship to either one or more variables. As the number of variables increases, identifying potential outliers using graphical representations becomes more difficult. In addition, special procedures are required to test for multivariate outliers. Details of these procedures are beyond the scope of this guidance. However, procedures for testing for multivariate outliers are contained in the software package Scout developed by the EPA's Environmental Monitoring Systems Laboratory in Las Vegas, Nevada (EMSL-LV) and statistical textbooks on multivariate analysis.

4.5 TESTS FOR DISPERSIONS

Many statistical tests make assumptions on the dispersion (as measured by variance) of data; this section considers some of the most commonly used statistical tests for variance assumptions. Section 4.5.1 contains the methodology for constructing a confidence interval for a single variance estimate from a sample. Section 4.5.2 deals with the equality of two variances, a key assumption for the validity of a two-sample t-test. Section 4.5.3 describes Bartlett's test and section 4.5.4 describes Levene's test. These two tests verify the assumption that two or more variances are equal, a requirement for a standard two-sample t-test, for example. The analyst should be aware that many statistical tests only require the assumption of approximate equality and that many of these tests remain valid unless gross inequality in variances is determined.

4.5.1 Confidence Intervals for a Single Variance

This section discusses confidence intervals for a single variance or standard deviation for analysts interested in the precision of variance estimates. This information may be necessary for performing a sensitivity analysis of the statistical test or analysis method. The method described in Box 4.5-1 can be used to find a two-sided $100(1-\alpha)\%$ confidence interval. The upper end point of a two-sided $100(1-\alpha)\%$ confidence *interval* is a $100(1-\alpha/2)\%$ upper confidence *limit*, and the lower end point of a two-sided $100(1-\alpha)\%$ confidence *interval* is a $100(1-\alpha/2)\%$ lower confidence *limit*. For example, the upper end point of a 90% confidence interval is a 95% upper confidence limit and the lower end point is a 95% lower confidence limit. Since the standard deviation is the square root of the variance, a confidence interval for the variance can be converted to a confidence interval for the standard deviation by taking the square roots of the endpoints of the interval. This confidence interval assumes that the data constitute a random sample from a normally distributed population and can be highly sensitive to outliers and to departures from normality.

4.5.2 The F-Test for the Equality of Two Variances

An F-test may be used to test whether the true underlying variances of two populations are equal. Usually the F-test is employed as a preliminary test, before conducting the two-sample t-test for the equality of two means. The assumptions underlying the F-test are that the two samples are independent random samples from two underlying normal populations. The F-test for equality of variances is highly sensitive to departures from normality. Directions for implementing an F-test with an example are given in Box 4.5-2.

4.5.3 Bartlett's Test for the Equality of Two or More Variances

Bartlett's test is a means of testing whether two or more population variances of normal distributions are equal. In the case of only two variances, Bartlett's test is equivalent to the F-test. Often in practice unequal variances and non-normality occur together and Bartlett's test is itself sensitive to departures from normality. With long-tailed distributions, the test too often rejects equality (homogeneity) of the variances.

Bartlett's test requires the calculation of the variance for each sample, then calculation of a statistic associated with the logarithm of these variances. This statistic is compared to tables and if it exceeds the tabulated value, the conclusion is that the variances differ as a complete set. It does *not* mean that one is significantly different from the others, nor that one or more are larger (smaller) than the rest. It simply implies the variances are unequal as a group. Directions for Bartlett's test are given in Box 4.5-3 and an example is given in Box 4.5-4.

Box 4.5-1: Directions for Constructing Confidence Intervals and Confidence Limits for the Sample Variance and Standard Deviation with an Example

Directions: Let $X_1, X_2, \ldots, X_n$ represent the n data points.

STEP 1: Calculate the sample variance s^2 (section 2.2.3).

STEP 2: For a 100(1-α)% two-sided confidence interval, use Table A-8 of Appendix A to find the cutoffs L and U such that $L = \chi^2_{\alpha/2}$ and $U = \chi^2_{(1-\alpha/2)}$ with (n-1) degrees of freedom (*dof*).

STEP 3: A 100(1-α)% confidence interval for the true underlying variance is: $\frac{(n-1)s^2}{L}$ *to* $\frac{(n-1)s^2}{U}$.

A 100(1-α)% confidence interval for the true standard deviation is: $\sqrt{\frac{(n-1)s^2}{L}}$ *to* $\sqrt{\frac{(n-1)s^2}{U}}$.

Example: Ten samples were analyzed for lead: 46.4, 46.1, 45.8, 47, 46.1, 45.9, 45.8, 46.9, 45.2, 46 ppb.

STEP 1: Using section 2.2.3, $s^2 = 0.286$.

STEP 2: Using Table A-8 of Appendix A and 9 *dof*, $L = \chi^2_{.05/2} = \chi^2_{.025} = 19.02$ and $U = \chi^2_{(1-.05/2)} = \chi^2_{.975} = 2.70$.

STEP 3: A 95% confidence interval for the variance is: $\frac{(10-1)0.286}{19.02}$ *to* $\frac{(10-1)0.286}{2.70}$ or 0.14 to 0.95.

A 95% confidence interval for the standard deviation is: $\sqrt{0.14} = .374$ *to* $\sqrt{0.95} = .975$.

Box 4.5-2: Directions for Calculating an F-Test to Compare Two Variances with an Example

Directions: Let $X_1, X_2, \ldots, X_m$ represent the m data points from population 1 and $Y_1, Y_2, \ldots, Y_n$ represent the n data points from population 2. To perform an F-test, proceed as follows.

STEP 1: Calculate the sample variances s_X^2 (for the X's) and s_Y^2 (for the Y's) (section 2.2.3).

STEP 2: Calculate the variance ratios $F_X = s_X^2/s_Y^2$ and $F_Y = s_Y^2/s_X^2$. Let F equal the larger of these two values. If $F = F_x$, then let k = m - 1 and q = n - 1. If $F = F_y$, then let k = n -1 and q = m - 1.

STEP 3: Using Table A-9 of Appendix A of the F distribution, find the cutoff $U = f_{1-\alpha/2}(k, q)$. If F > U, conclude that the variances of the two populations are not the same.

Example: Manganese concentrations were collected from 2 wells. The data are Well X: 50, 73, 244, and 202 ppm; and Well Y: 272, 171, 32, 250, and 53 ppm. An F-test will be used to determine if the variances of the two wells are equal.

STEP 1: For Well X, $s_X^2 = 9076$. For Well Y, $s_Y^2 = 12125$.

STEP 2: $F_X = s_X^2/s_Y^2 = 9076 / 12455 = 0.749$. $F_Y = s_Y^2/s_X^2 = 12445 / 9076 = 1.334$. Since, $F_Y > F_X$, $F = F_Y = 1.334$, k = 5 -1 = 4 and q = 4 - 1 = 3.

STEP 3: Using Table A-9 of Appendix A of the F distribution with α = 0.05, $L = f_{1-.05/2}(4, 3) = 15.1$. Since 1.334 < 15.1, there is no evidence that the variability of the two wells is different.

Box 4.5-3: Directions for Bartlett's Test

Consider k groups with a sample size of n_i for each group. Let N represent the total number of samples, i.e., let $N = n_1 + n_2 + \ldots + n_k$. For example, consider two wells where 4 samples have been taken from well 1 and 3 samples have been taken from well 2. In this case, k = 2, $n_1 = 4$, $n_2 = 3$, and N = 4 + 3 = 7.

STEP 1: For each of the k groups, calculate the sample variances, s_i^2 (section 2.2.3).

STEP 2: Compute the pooled variance across groups: $s_p^2 = \frac{1}{(N-k)} \sum_{i=1}^{k} (n_i - 1) s_i^2$

STEP 3: Compute the test statistic: $TS = (N - k) \ln(s_p^2) - \sum_{i=1}^{k} (n_i - 1) \ln(s_i^2)$

where "ln" stands for natural logarithms.

STEP 4: Using a chi-squared table (Table A-8 of Appendix A), find the critical value for χ^2 with (k-1) degrees of freedom at a predetermined significance level. For example, for a significance level of 5% and 5 degrees of freedom, $\chi^2 = 11.1$. If the calculated value (*TS*) is greater than the tabulated value, conclude that the variances are not equal at that significance level.

Box 4.5-4: An Example of Bartlett's Test

Manganese concentrations were collected from 6 wells over a 4 month period. The data are shown in the following table. Before analyzing the data, it is important to determine if the variances of the six wells are equal. Bartlett's test will be used to make this determination.

STEP 1: For each of the 6 wells, the sample means and variances were calculated. These are shown in the bottom rows of the table below.

Sampling Date	Well 1	Well 2	Well 3	Well 4	Well 5	Well 6
January 1	50		272			
February 1	73		171			68
March 1	244	46	32	34	48	991
April 1	202	77	53	3940	54	54
n_i (N=17)	4	2	4	2	2	3
$\bar{x}_i$	142.25	61.50	132	1987	51.00	371.00
s_i^2	9076.37	480.49	12455	7628243	17.98	288348

STEP 2: $s_p^2 = \frac{1}{(N-k)} \sum_{i=1}^{k} (n_i - 1) s_i^2 = \frac{1}{(17-6)} [(4-1)9076 + \ldots + (3-1)576696] = 751837.27$

STEP 3: $TS = (17-6) \ln(751837.27) - [\,(4-1)\ln(9076) + \ldots + (3-1)\ln(288348)\,] = 43.16$

STEP 4: The critical χ^2 value with 6 - 1 = 5 degrees of freedom at the 5% significance level is 11.1 (from Table A-8 of Appendix A). Since 43.16 is larger than 11.1, it is concluded that the six variances ($s_1^2, \ldots, s_6^2$) are not homogeneous at the 5% significance level.

4.5.4 Levene's Test for the Equality of Two or More Variances

Levene's test provides an alternative to Bartlett's test for homogeneity of variance (testing for differences among the dispersions of several groups). Levene's test is less sensitive to departures from normality than Bartlett's test and has greater power than Bartlett's for non-normal data. In addition, Levene's test has power nearly as great as Bartlett's test for normally distributed data. However, Levene's test is more difficult to apply than Bartlett's test since it involves applying an analysis of variance (ANOVA) to the absolute deviations from the group means. Directions and an example of Levene's test are contained in Box 4.5-5 and Box 4.5-6, respectively.

Box 4.5-5: Directions for Levene's Test

Consider k groups with a sample size of n_i for the ith group. Let N represent the total number of samples, i.e., let $N = n_1 + n_2 + \ldots + n_k$. For example, consider two wells where 4 samples have been taken from well 1 and 3 samples have been taken from well 2. In this case, k = 2, $n_1 = 4$, $n_2 = 3$, and N = 4 + 3 = 7.

STEP 1: For each of the k groups, calculate the group mean, $\bar{x}_i$ (section 2.2.2), i.e., calculate:

$$\bar{X}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j}, \quad \bar{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}, \quad \ldots, \quad \bar{X}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} x_{kj}.$$

STEP 2: Compute the absolute residuals $z_{ij} = \left|X_{ij} - \bar{X}_i\right|$ where X_{ij} represents the jth value of the ith group.

For each of the k groups, calculate the means, $\bar{z}_i$, of these residuals, i.e., calculate:

$$\bar{z}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} z_{1j}, \quad \bar{z}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} z_{2j}, \quad \ldots, \quad \bar{z}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} z_{kj}.$$

Also calculate the overall mean residual as $\bar{z} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij} = \frac{1}{N}\sum_{i=1}^{k} n_i \bar{z}_i$.

STEP 3: Compute the following sums of squares for the absolute residuals:

$$SS_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij}^2 - \frac{\bar{z}}{N}, \quad SS_{GROUPS} = \sum_{i=1}^{k}\frac{\bar{z}_i^2}{n_i} - \frac{\bar{z}}{N},$$ and $SS_{ERROR} = SS_{TOTAL} - SS_{GROUPS}$.

STEP 4: Compute $f = \dfrac{SS_{GROUPS}/(k-1)}{SS_{ERROR}/(N-k)}$

STEP 5: Using Table A-9 of Appendix A, find the critical value of the F-distribution with (k-1) numerator degrees of freedom, (N-k) denominator degrees of freedom, and a desired level of significance (α). For example, if $\alpha = 0.05$, the numerator degrees of freedom is 5, and the denominator degrees of freedom is 18, then using Table A-9, F = 2.77. If f is greater than F, reject the assumptions of equal variances.

Box 4.5-6: An Example of Levene's Test

Four months of data on arsenic concentration were collected from six wells at a Superfund site. This data set is shown in the table below. Before analyzing this data, it is important to determine if the variances of the six wells are equal. Levene's test will be used to make this determination.

STEP 1: The group mean for each well ($\bar{x}_i$) is shown in the last row of the table below.

Arsenic Concentration (ppm)						
Month	Well 1	Well 2	Well 3	Well 4	Well 5	Well 6
1	22.90	2.00	2.0	7.84	24.90	0.34
2	3.09	1.25	109.4	9.30	1.30	4.78
3	35.70	7.80	4.5	25.90	0.75	2.85
4	4.18	52.00	2.5	2.00	27.00	1.20
Group Means	$\bar{x}_1$=16.47	$\bar{x}_2$=15.76	$\bar{x}_3$=29.6	$\bar{x}_4$=11.26	$\bar{x}_5$=13.49	$\bar{x}_6$=2.29

STEP 2: To compute the absolute residuals z_{ij} in each well, the value 16.47 will be subtracted from Well 1 data, 15.76 from Well 2 data, 29.6 from Well 3 data, 11.26 from Well 4 data, 13.49 from Well 5 data, and 2.29 from Well 6 data. The resulting values are shown in the following table with the new well means ($\bar{z}_i$) and the total mean $\bar{z}$.

Residual Arsenic Concentration (ppm)						
Month	Well 1	Well 2	Well 3	Well 4	Well 5	Well 6
1	6.43	13.76	27.6	3.42	11.41	1.95
2	13.38	14.51	79.8	1.96	12.19	2.49
3	19.23	7.96	25.1	14.64	12.74	0.56
4	12.29	36.24	27.1	9.26	13.51	1.09
Residual Means	$\bar{z}_1$=12.83	$\bar{z}_2$=18.12	$\bar{z}_3$=39.9	$\bar{z}_4$=7.32	$\bar{z}_5$=12.46	$\bar{z}_6$=1.52
Total Residual Mean $\bar{z}$ = (1/6)(12.83 + 18.12 + 39.9 + 7.32 + 12.46 + 1.52) = 15.36						

STEP 3: The sum of squares are: SS_{TOTAL} = 6300.89, SS_{WELLS} = 3522.90, and SS_{ERROR} = 2777.99.

STEP 4: $$f = \frac{SS_{WELLS}/(k-1)}{SS_{ERROR}/(N-k)} = \frac{3522.9/(6-1)}{2777.99/(24-6)} = 4.56$$

STEP 5: Using Table A-9 of Appendix A, the F statistic for 5 and 18 degrees of freedom with α = 0.05 is 2.77. Since f=4.56 exceeds $F_{.05}$=2.77, the assumption of equal variances should be rejected.

4.6 TRANSFORMATIONS

Most statistical tests and procedures contain assumptions about the data to which they will be applied. For example, some common assumptions are that the data are normally distributed; variance components of a statistical model are additive; two independent data sets have equal variance; and a data set has no trends over time or space. If the data do not satisfy such assumptions, then the results of a statistical procedure or test may be biased or incorrect. Fortunately, data that do not satisfy statistical assumptions may often be converted or transformed mathematically into a form that allows standard statistical tests to perform adequately.

4.6.1 Types of Data Transformations

Any mathematical function that is applied to every point in a data set is called a transformation. Some commonly used transformations include:

Logarithmic (Log X or Ln X): This transformation may be used when the original measurement data follow a lognormal distribution or when the variance at each level of the data is proportional to the square of the mean of the data points at that level. For example, if the variance of data collected around 50 ppm is approximately 250, but the variance of data collected around 100 ppm is approximately 1000, then a logarithmic transformation may be useful. This situation is often characterized by having a constant coefficient of variation (ratio of standard deviation to mean) over all possible data values.

The logarithmic base (for example, either natural or base 10) needs to be consistent throughout the analysis. If some of the original values are zero, it is customary to add a small quantity to make the data value non-zero as the logarithm of zero does not exist. The size of the small quantity depends on the magnitude of the non-zero data and the consequences of potentially erroneous inference from the resulting transformed data. As a working point, a value of one tenth the smallest non-zero value could be selected. It does not matter whether a natural (ln) or base 10 (log) transformation is used because the two transformations are related by the expression ln(X) = 2.303 log(X). Directions for applying a logarithmic transformation with an example are given in Box 4.6-1.

Square Root ($\sqrt{X}$): This transformation may be used when dealing with small whole numbers, such as bacteriological counts, or the occurrence of rare events, such as violations of a standard over the course of a year. The underlying assumption is that the original data follow a Poisson-like distribution in which case the mean and variance of the data are equal. It should be noted that the square root transformation overcorrects when very small values and zeros appear in the original data. In these cases, $\sqrt{X+1}$ is often used as a transformation.

Inverse Sine (Arcsine X): This transformation may be used for binomial proportions based on count data to achieve stability in variance. The resulting transformed data are expressed in radians (angular degrees). Special tables must be used to transform the proportions into degrees.

Box-Cox Transformations: This transformation is a complex power transformation that takes the original data and raises each data observation to the power lambda (λ). A logarithmic transformation is a special case of the Box-Cox transformation. The rationale is to find λ such that the transformed data have the best possible additive model for the variance structure, the errors are normally

distributed, and the variance is as constant as possible over all possible concentration values. The Maximum Likelihood technique is used to find λ such that the residual error from fitting the theorized model is minimized. In practice, the exact value of λ is often rounded to a convenient value for ease in interpretation (for example, $\lambda = -1.1$ would be rounded to -1 as it would then have the interpretation of a reciprocal transform). One of the drawbacks of the Box-Cox transformation is the difficulty in physically interpreting the transformed data.

4.6.2 Reasons for Data Transformations

By transforming the data, assumptions that are not satisfied in the original data can be satisfied by the transformed data. For instance, a right-skewed distribution can be transformed to be approximately Gaussian (normal) by using a logarithmic or square-root transformation. Then the normal-theory procedures can be applied to the transformed data. If data are lognormally distributed, then apply procedures to logarithms of the data. However, selecting the correct transformation may be difficult. If standard transformations do not apply, it is suggested that the data user consult a statistician.

Another important use of transformations is in the interpretation of data collected under conditions leading to an Analysis of Variance (ANOVA). Some of the key assumptions needed for analysis (for example, additivity of variance components) may only be satisfied if the data are transformed suitably. The selection of a suitable transformation depends on the structure of the data collection design; however, the interpretation of the transformed data remains an issue.

While transformations are useful for dealing with data that do not satisfy statistical assumptions, they can also be used for various other purposes. For example, transformations are useful for consolidating data that may be spread out or that have several extreme values. In addition, transformations can be used to derive a linear relationship between two variables, so that linear regression analysis can be applied. They can also be used to efficiently estimate quantities such as the mean and variance of a lognormal distribution. Transformations may also make the analysis of data easier by changing the scale into one that is more familiar or easier to work with.

Once the data have been transformed, all statistical analysis must be performed on the transformed data. No attempt should be made to transform the data back to the original form because this can lead to biased estimates. For example, estimating quantities such as means, variances, confidence limits, and regression coefficients in the transformed scale typically leads to biased estimates when transformed back into original scale. However, it may be difficult to understand or apply results of statistical analysis expressed in the transformed scale. Therefore, if the transformed data do not give noticeable benefits to the analysis, it is better to use the original data. There is no point in working with transformed data unless it adds value to the analysis.

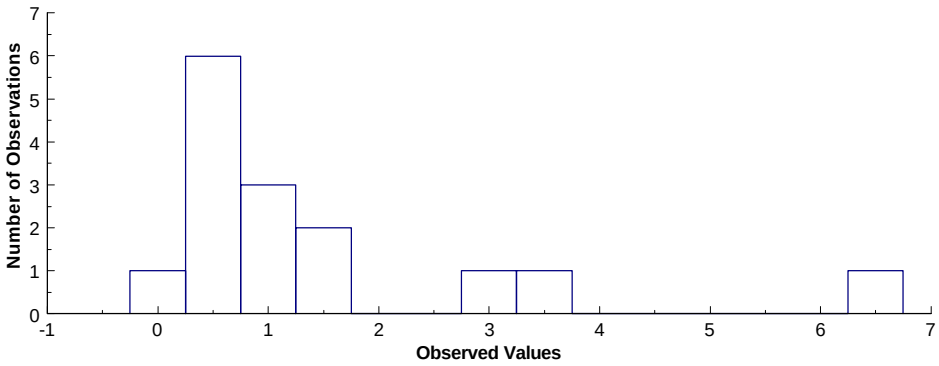
Box 4.6-1: Directions for Transforming Data and an Example

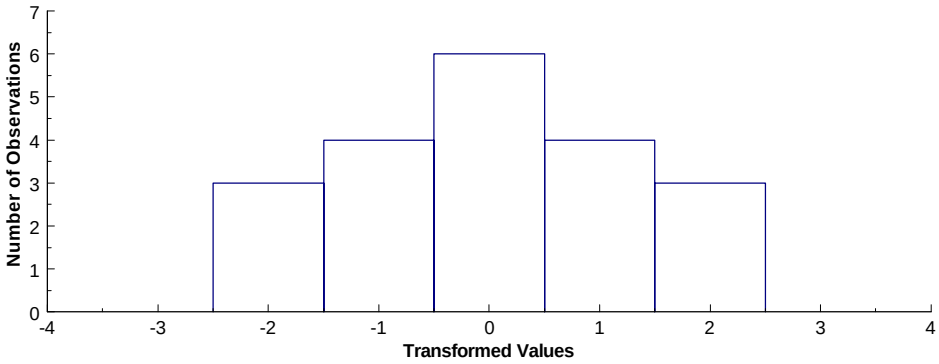
Let $X_1, X_2, \ldots, X_n$ represent the n data points. To apply a transformation, simply apply the transforming function to each data point. When a transformation is implemented to make the data satisfy some statistical assumption, it will need to be verified that the transformed data satisfy this assumption.

Example: Transforming Lognormal Data

A logarithmic transformation is particularly useful for pollution data. Pollution data are often skewed, thus the log-transformed data will tend to be symmetric. Consider the data set shown below with 15 data points. The frequency plot of this data (below) shows that the data are possibly lognormally distributed. If any analysis performed with this data assumes normality, then the data may be logarithmically transformed to achieve normality. The transformed data are shown in column 2. A frequency plot of the transformed data (below) shows that the transformed data appear to be normally distributed.

Observed X	→	Transformed ln(X)	Observed X	→	Transformed ln(X)
0.22	→	-1.51	0.47	→	-0.76
3.48	→	1.25	0.67	→	-0.40
6.67	→	1.90	0.75	→	-0.29
2.53	→	0.93	0.60	→	-0.51
1.11	→	0.10	0.99	→	-0.01
0.33	→	-1.11	0.90	→	-0.11
1.64	→	0.50	0.26	→	-1.35
1.37	→	0.31			





4.7 VALUES BELOW DETECTION LIMITS

Data generated from chemical analysis may fall below the detection limit (DL) of the analytical procedure. These measurement data are generally described as not detected, or nondetects, (rather than as zero or not present) and the appropriate limit of detection is usually reported. In cases where measurement data are described as not detected, the concentration of the chemical is unknown although it lies somewhere between zero and the detection limit. Data that includes both detected and non-detected results are called censored data in the statistical literature.

There are a variety of ways to evaluate data that include values below the detection limit. However, there are no general procedures that are applicable in all cases. Some general guidelines are presented in Table 4.7-1. Although these guidelines are usually adequate, they should be implemented cautiously.

Percentage of Nondetects	Section	Statistical Analysis Method
< 15%	4.7.1	Replace nondetects with DL/2, DL, or a very small number.
15% - 50%	4.7.2	Trimmed mean, Cohen's adjustment, Winsorized mean and standard deviation.
> 50% - 90%	4.7.3	Use tests for proportions (section 3.2.2)

Table 4.7-1. Guidelines for Analyzing Data with Nondetects

All of the suggested procedures for analyzing data with nondetects depend on the amount of data below the detection limit. For relatively small amounts below detection limit values, replacing the nondetects with a small number and proceeding with the usual analysis may be satisfactory. For moderate amounts of data below the detection limit, a more detailed adjustment is appropriate. In situations where relatively large amounts of data below the detection limit exist, one may need only to consider whether the chemical was detected as above some level or not. The interpretation of small, moderate, and large amounts of data below the DL is subjective. Table 4.7-1 provides percentages to assist the user in evaluating their particular situation. However, it should be recognized that these percentages are not hard and fast rules, but should be based on judgement.

In addition to the percentage of samples below the detection limit, sample size influences which procedures should be used to evaluate the data. For example, the case where 1 sample out of 4 is not detected should be treated differently from the case where 25 samples out of 100 are not detected. Therefore, this guidance suggests that the data analyst consult a statistician for the most appropriate way to evaluate data containing values below the detection level.

4.7.1 Less than 15% Nondetects - Substitution Methods

If a small proportion of the observations are not detected, these may be replaced with a small number, usually the detection limit divided by 2 (DL/2), and the usual analysis performed. As a guideline, if 15% or fewer of the values are not detected, replace them with the method detection limit divided by two and proceed with the appropriate analysis using these modified values. If simple substitution of values below the detection limit is proposed when more than 15% of the values are reported as not detected, consider using nonparametric methods or a test of proportions to analyze the data. If a more accurate method is to be considered, see Cohen's Method (section 4.7.2.1).

4.7.2 Between 15-50% Nondetects

4.7.2.1 Cohen's Method

Cohen's method provides adjusted estimates of the sample mean and standard deviation that accounts for data below the detection level. The adjusted estimates are based on the statistical technique of maximum likelihood estimation of the mean and variance so that the fact that the nondetects are below the limit of detection but may not be zero is accounted for. The adjusted mean and standard deviation can then be used in the parametric tests described in Chapter 3 (e.g., the one sample t-test of section 3.2.1.1). However, if more than 50% of the observations are not detected, Cohen's method should not be used. In addition, this method requires that the data without the nondetects be normally distributed and the detection limit is always the same. Directions for Cohen's method are contained in Box 4.7-1; an example is given in Box 4.7-2.

Box 4.7-1: Directions for Cohen's Method

Let $X_1, X_2, \ldots, X_n$ represent the n data points with the first m values representing the data points above the detection limit (DL). Thus, there are (n-m) data points are below the DL.

STEP 1: Compute the sample mean $\bar{x}_d$ from the data above the detection limit: $\bar{X}_d = \frac{1}{m}\sum_{i=1}^{m} X_i$

STEP 2: Compute the sample variance s_d^2 from the data above the detection limit:

$$s_d^2 = \frac{\sum_{i=1}^{m} X_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} X_i\right)^2}{m-1}$$

STEP 3: Compute $h = \frac{(n-m)}{n}$ and $\gamma = \frac{s_d^2}{(\bar{X}_d - DL)^2}$

STEP 4: Use h and γ in Table A-10 of Appendix A to determine $\hat{\lambda}$. For example, if h = 0.4 and γ = 0.30, then $\hat{\lambda}$ = 0.6713. If the exact value of h and γ do not appear in the table, use double linear interpolation (Box 4.7-3) to estimate $\hat{\lambda}$.

STEP 5: Estimate the corrected sample mean, $\bar{x}$, and sample variance, s^2, to account for the data below the detection limit, as follows: $\bar{X} = \bar{X}_d - \hat{\lambda}(\bar{X}_d - DL)$ and $s^2 = s_d^2 + \hat{\lambda}(\bar{X}_d - DL)^2$.

Box 4.7-2: An Example of Cohen's Method

Sulfate concentrations were measured for 24 data points. The detection limit was 1,450 mg/L and 3 of the 24 values were below the detection level. The 24 values are 1850, 1760, < 1450 (ND), 1710, 1575, 1475, 1780, 1790, 1780, < 1450 (ND), 1790, 1800, < 1450 (ND), 1800, 1840, 1820, 1860, 1780, 1760, 1800, 1900, 1770, 1790, 1780 mg/L. Cohen's Method will be used to adjust the sample mean for use in a t-test to determine if the mean is greater than 1600 mg/L.

STEP 1: The sample mean of the m = 21 values above the detection level is $\bar{X}_d = 1771.9$

STEP 2: The sample variance of the 21 quantified values is $s_d^2 = 8593.69$.

STEP 3: $h = (24-21)/24 = 0.125$ and $\gamma = 8593.69/(1771.9-1450)^2 = 0.083$

STEP 4: Table A-10 of Appendix A was used for h = 0.125 and $\gamma = 0.083$ to find the value of $\hat{\lambda}$. Since the table does not contain these entries exactly, double linear interpolation was used to estimate $\hat{\lambda} = 0.14986$ (see Box 4.7-3).

STEP 5: The corrected sample mean and standard deviation are then estimated as follows:

$$\bar{X} = 1771.9 - 0.14986(1771.9-1450) = 1723.66 \text{ and}$$

$$s^2 = 8593.69 + 0.14986(1771.9-1450)^2 = 24122.12$$

Box 4.7-3: Double Linear Interpolation

The details of the double linear interpolation are provided to assist in the use of Table A-10 of Appendix A. The desired value for $\hat{\lambda}$ corresponds to $\gamma = 0.083$ and, h = 0.125 from Box 4.7-2, Step 3. The values from Table A-10 for interpolatation are:

γ	h = 0.10	h = 0.15
0.05	0.11431	0.17935
0.10	0.11804	0.18479

There are 0.05 units between 0.10 and 0.15 on the h-scale and 0.025 units between 0.10 and 0.125. Therefore, the value of interest lies (0.025/0.05)100% = 50% of the distance along the interval between 0.10 and 0.15. To linearly interpolate between tabulated values on the h axis for $\gamma = 0.05$, the range between the values must be calculated, 0.17935 - 0.11431 = 0.06504; the value that is 50% of the distance along the range must be computed, 0.06504 x 0.50 = 0.03252; and then that value must be added to the lower point on the tabulated values, 0.11431 + 0.03252 = 0.14683. Similarly for $\gamma = 0.10$, 0.18479 - 0.11804 = 0.06675, 0.06675 x 0.50 = 0.033375, and 0.11804 + 0.033375 = 0.151415.

On the γ-axis there are 0.033 units between 0.05 and 0.083 and there are 0.05 units between 0.05 and 0.10. The value of interest (0.083) lies (0.033/0.05 x 100) = 66% of the distance along the interval between 0.05 and 0.10, so 0.141415 - 0.14683 = 0.004585, 0.004585 * 0.66 = 0.0030261. Therefore,

$$\hat{\lambda} = 0.14683 + 0.0030261 = 0.14986.$$

4.7.2.2 Trimmed Mean

Trimming discards the data in the tails of a data set in order to develop an unbiased estimate of the population mean. For environmental data, nondetects usually occur in the left tail of the data so trimming the data can be used to adjust the data set to account for nondetects when estimating a mean. Developing a 100p% trimmed mean involves trimming p% of the data in both the lower and the upper tail. Note that p must be between 0 and .5 since p represents the portion deleted in both the upper and the lower tail. After np of the largest values and np of the smallest values are trimmed, there are n(1-2p) data values remaining. Therefore, the proportion trimmed is dependent on the total sample size (n) since a reasonable amount of samples must remain for analysis. For approximately symmetric distributions, a 25% trimmed mean (the midmean) is a good estimator of the population mean. However, environmental data are often skewed (non-symmetric) and in these cases a 15% trimmed mean performance may be a good estimator of the population mean. It is also possible to trim the data only to replace the nondetects. For example, if 3% of the data are below the detection limit, a 3% trimmed mean could be used to estimate the population mean. Directions for developing a trimmed mean are contained in Box 4.7-4 and an example is given in Box 4.7-5. A trimmed variance is rarely calculated and is of limited use.

Box 4.7-4: Directions for Developing a Trimmed Mean

Let $X_1, X_2, \ldots, X_n$ represent the n data points. To develop a 100p% trimmed mean ($0 < p < 0.5$):

STEP 1: Let t represent the integer part of the product np. For example, if p = .25 and n = 17, np = (.25)(17) = 4.25, so t = 4.

STEP 2: Delete the t smallest values of the data set and the t largest values of the data set.

STEP 3: Compute the arithmetic mean of the remaining n - 2t values: $\bar{X} = \frac{1}{n-2t}\sum_{i=1}^{n-2t} X_i$

This value is the estimate of the population mean.

Box 4.7-5: An Example of the Trimmed Mean

Sulfate concentrations were measured for 24 data points. The detection limit was 1,450 mg/L and 3 of the 24 values were below this limit. The 24 values listed in order from smallest to largest are: < 1450 (ND), < 1450 (ND), < 1450 (ND), 1475, 1575, 1710, 1760, 1760, 1770, 1780, 1780, 1780, 1780, 1790, 1790, 1790, 1800, 1800, 1800, 1820, 1840, 1850, 1860, 1900 mg/L. A 15% trimmed mean will be used to develop an estimate of the population mean that accounts for the 3 nondetects.

STEP 1: Since np = (24)(.15) = 3.6, t = 3.

STEP 2: The 3 smallest values of the data set and the 3 largest values of the data set were deleted. The new data set is: 1475, 1575, 1710, 1760, 1760, 1770, 1780, 1780, 1780, 1780, 1790, 1790, 1790, 1800, 1800, 1800, 1820, 1840 mg/L.

STEP 3: Compute the arithmetic mean of the remaining n-2t values:

$$\bar{X} = \frac{1}{24-(2)(3)}(1475 + \ldots + 1840) = 1755.56$$

Therefore, the 15% trimmed mean is 1755.56 mg/L, which is an estimate of the population mean.

4.7.2.3 Winsorized Mean and Standard Deviation

Winsorizing replaces data in the tails of a data set with the next most extreme data value. For environmental data, nondetects usually occur in the left tail of the data. Therefore, winsorizing can be used to adjust the data set to account for nondetects. The mean and standard deviation can then be computed on the new data set. Directions for winsorizing data (and revising the sample size) are contained in Box 4.7-6 and an example is given in Box 4.7-7.

Box 4.7-6: Directions for Developing a Winsorized Mean and Standard Deviation

Let $X_1, X_2, \ldots, X_n$ represent the n data points and m represent the number of data points above the detection limit (DL), and hence n-m below the DL.

STEP 1: List the data in order from smallest to largest, including nondetects. Label these points $X_{(1)}$, $X_{(2)}, \ldots, X_{(n)}$ (so that $X_{(1)}$ is the smallest, $X_{(2)}$ is the second smallest, and $X_{(n)}$ is the largest).

STEP 2: Replace the n-m nondetects with $X_{(m+1)}$ and replace the n-m largest values with $X_{(n-m)}$.

STEP 3: Using the revised data set, compute the sample mean, $\bar{x}$, and the sample standard deviation, s:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad s = \sqrt{\frac{(\sum_{i=1}^{n} X_i^2) - n\bar{X}^2}{n-1}}$$

STEP 4: The Winsorized mean $\bar{x}_w$ is equal to $\bar{x}$. The Winsorized standard deviation is $s_w = \frac{s(n-1)}{(2m-n-1)}$.

Box 4.7-7: An Example of a Winsorized Mean and Standard Deviation

Sulfate concentrations were measured for 24 data points. The detection limit was 1,450 mg/L and 3 of the 24 values were below the detection level. The 24 values listed in order from smallest to largest are: < 1450 (ND), < 1450 (ND), < 1450 (ND), 1475, 1575, 1710, 1760, 1760, 1770, 1780, 1780, 1780, 1780, 1790, 1790, 1790, 1800, 1800, 1800, 1820, 1840, 1850, 1860, 1900 mg/L.

STEP 1: The data above are already listed from smallest to largest. There are n=24 samples, 21 above DL, and n-m=3 nondetects.

STEP 2: The 3 nondetects were replaced with $X_{(4)}$ and the 3 largest values were replaced with $X_{(21)}$. The resulting data set is: 1475, 1475, 1475, 1475, 1575, 1710, 1760, 1760, 1770, 1780, 1780, 1780, 1780, 1790, 1790, 1790, 1800, 1800, 1800, 1820, 1840, 1840, 1840, 1840 mg/L

STEP 3: For the new data set, $\bar{x}$ = 1731 mg/L and s = 128.52 mg/L.

STEP 4: The Winsorized mean $\bar{x}_w$ = 1731 mg/L. The Winsorized sample standard deviation is:

$$s_w = \frac{128.52(24-1)}{2(21)-24-1} = 173.88$$

4.7.3 Greater than 50% Nondetects - Test of Proportions

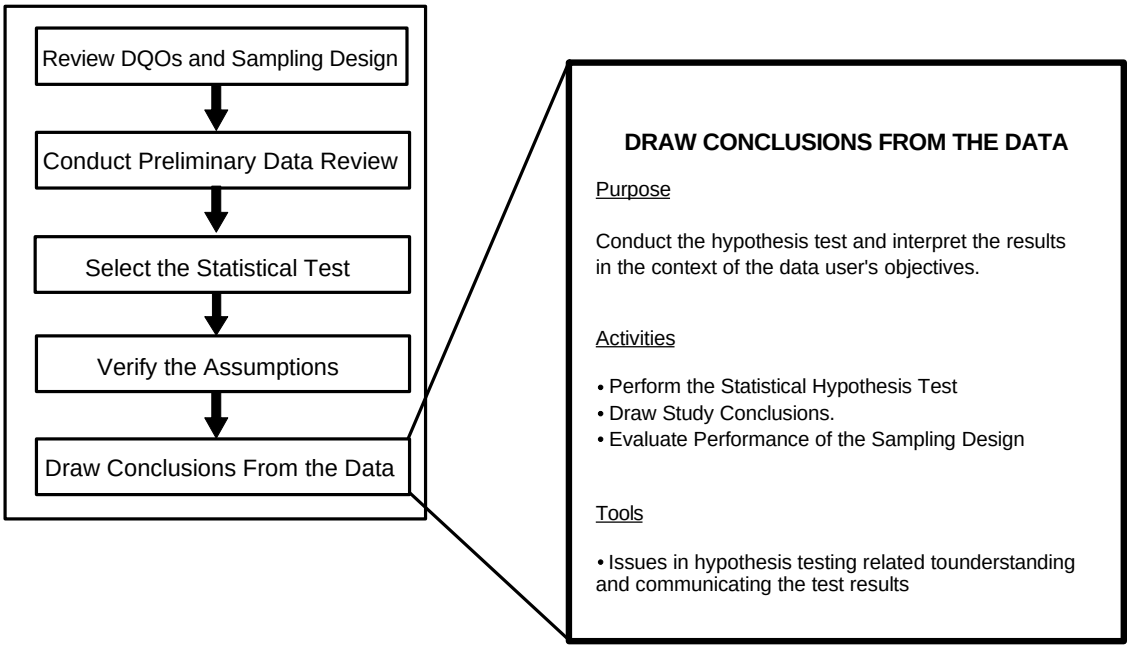
If more than 50% of the data are below the detection limit but at least 10% of the observations are quantified, tests of proportions may be used to test hypotheses using the data. Thus, if the parameter of interest is a mean, consider switching the parameter of interest to some percentile greater than the percent of data below the detection limit. For example, if 67% of the data are below the DL, consider switching the parameter of interest to the 75 th percentile. Then the method described in 3.2.2 can be applied to test the hypothesis concerning the 75 th percentile. It is important to note that the tests of proportions may not be applicable for composite samples. In this case, the data analyst should consult a statistician before proceeding with analysis.

If very few quantified values are found, a method based on the Poisson distribution may be used as an alternative approach. However, with a large proportion of nondetects in the data, the data analyst should consult with a statistician before proceeding with analysis.

CHAPTER 5

STEP 5: DRAW CONCLUSIONS FROM THE DATA

THE DATA QUALITY ASSESSMENT PROCESS



Step 5: Draw Conclusions from the Data

- Perform the calculations for the statistical hypothesis test.
  - Perform the calculations and document them clearly.
  - If anomalies or outliers are present in the data set, perform the calculations with and without the questionable data.
- Evaluate the statistical test results and draw conclusions.
  - If the null hypothesis is rejected, then draw the conclusions and document the analysis.
  - If the null hypothesis is not rejected, verify whether the tolerable limits on false negative decision errors have been satisfied. If so, draw conclusions and document the analysis; if not, determine corrective actions, if any.
  - Interpret the results of the test.
- Evaluate the performance of the sampling design if the design is to be used again.
  - Evaluate the statistical power of the design over the full range of parameter values; consult a statistician as necessary.

STEP 5: DRAW CONCLUSIONS FROM THE DATA

		Page
5.1	OVERVIEW AND ACTIVITIES	5.1 - 1
5.1.1	Perform the Statistical Hypothesis Test	5.1 - 1
5.1.2	Draw Study Conclusions	5.1 - 1
5.1.3	Evaluate Performance of the Sampling Design	5.1 - 2
5.2	INTERPRETING AND COMMUNICATING THE TEST RESULTS	5.2 - 1
5.2.1	Interpretation of p-Values	5.2 - 1
5.2.2	"Accepting" vs. "Failing to Reject" the Null Hypothesis	5.2 - 1
5.2.3	Statistical Significance vs. Practical Significance	5.2 - 2
5.2.4	Impact of Bias on Test Results	5.2 - 2
5.2.5	Quantity vs. Quality of Data	5.2 - 5
5.2.6	"Proof of Safety" vs. "Proof of Hazard"	5.2 - 6

LIST OF FIGURES

Figure No.	Page
5.2-1. Illustration of Unbiased versus Biased Power Curves	5.2 - 5

LIST OF BOXES

Box No.	Page
Box 5.1-1: Checking Adequacy of Sample Size for a One-Sample t-Test	5.1 - 3
Box 5.1-2: Example of Power Calculations for the One-Sample Test of a Single Proportion	5.1 - 3
Box 5.2-1: Example of a Comparison of Two Variances which is Statistically but not Practically Significant	5.2 - 3
Box 5.2-2: Example of a Comparison of Two Biases	5.2 - 4

CHAPTER 5
STEP 5: DRAW CONCLUSIONS FROM THE DATA

5.1 OVERVIEW AND ACTIVITIES

In this final step of the DQA Process, the analyst performs the statistical hypothesis test and draws conclusions that address the data user's objectives. This step represents the culmination of the planning, implementation, and assessment phases of the data operations. The data user's planning objectives will have been reviewed (or developed retrospectively) and the sampling design examined in Step 1. Reports on the implementation of the sampling scheme will have been reviewed and a preliminary picture of the sampling results developed in Step 2. In light of the information gained in Step 2, the statistical test will have been selected in Step 3. To ensure that the chosen statistical methods are valid, the key underlying assumptions of the statistical test will have been verified in Step 4. Consequently, all of the activities conducted up to this point should ensure that the calculations performed on the data set and the conclusions drawn here in Step 5 address the data user's needs in a scientifically defensible manner. This chapter describes the main activities that should be conducted during this step. The actual procedures for implementing some commonly used statistical tests are described in Step 3, Select the Statistical Test.

5.1.1 Perform the Statistical Hypothesis Test

The goal of this activity is to conduct the statistical hypothesis test. Step-by-step directions for several commonly used statistical tests are described in Chapter 3. The calculations for the test should be clearly documented and easily verifiable. In addition, the documentation of the results of the test should be understandable so that the results can be communicated effectively to those who may hold a stake in the resulting decision. If computer software is used to perform the calculations, ensure that the procedures are adequately documented, particularly if algorithms have been developed and coded specifically for the project.

The analyst should always exercise best professional judgment when performing the calculations. For instance, if outliers or anomalies are present in the data set, the calculations should be performed both with and without the questionable data to see what effect they may have on the results.

5.1.2 Draw Study Conclusions

The goal of this activity is to translate the results of the statistical hypothesis test so that the data user may draw a conclusion from the data. The results of the statistical hypothesis test will be either:

(a) *reject the null hypothesis*, in which case the analyst is concerned about a possible false positive decision error; or

(b) *fail to reject the null hypothesis*, in which case the analyst is concerned about a possible false negative decision error.

In case (a), the data have provided the evidence needed to reject the null hypothesis, so the decision can be made with sufficient confidence and without further analysis. This is because the statistical test based on the classical hypothesis testing philosophy, which is the approach described in prior chapters, inherently controls the false positive decision error rate within the data user's tolerable limits, provided that the underlying assumptions of the test have been verified correctly.

In case (b), the data do not provide sufficient evidence to reject the null hypothesis, and the data must be analyzed further to determine whether the data user's tolerable limits on false negative decision errors have been satisfied. One of two possible conditions may prevail:

(1) The data do not support rejecting the null hypothesis and the false negative decision error limits were satisfied. In this case, the conclusion is drawn in favor of the null hypothesis, since the probability of committing a false negative decision error is believed to be sufficiently small in the context of the current study (see section 5.2).

(2) The data do not support rejecting the null hypothesis, and the false negative decision error limits were *not* satisfied. In this case, the statistical test was not powerful enough to satisfy the data user's performance criteria. The data user may choose to tolerate a higher false negative decision error rate than previously specified and draw the conclusion in favor of the null hypothesis, or instead take some form of corrective action, such as obtaining additional data before drawing a conclusion and making a decision.

When the test fails to reject the null hypothesis, the most thorough procedure for verifying whether the false negative decision error limits have been satisfied is to compute the estimated power of the statistical test, using the variability observed in the data. Computing the power of the statistical test across the full range of possible parameter values can be complicated and usually requires specialized software. Power calculations are also necessary for evaluating the performance of a sampling design. Thus, power calculations will be discussed further in section 5.1.3.

A simpler method can be used for checking the performance of the statistical test. Using an estimate of variance obtained from the actual data or upper 95% confidence limit on variance, the sample size required to satisfy the data user's objectives can be calculated retrospectively. If this theoretical sample size is less than or equal to the number of samples actually taken, then the test is sufficiently powerful. If the required number of samples is greater than the number actually collected, then additional samples would be required to satisfy the data user's performance criteria for the statistical test. An example of this method is contained in Box 5.1-1. The equations required to perform these calculations have been provided in the detailed step-by-step instructions for each hypothesis test procedure in Chapter 3.

5.1.3 Evaluate Performance of the Sampling Design

If the sampling design is to be used again, either in a later phase of the current study or in a similar study, the analyst will be interested in evaluating the overall performance of the design. To evaluate the sampling design, the analyst performs a statistical power analysis that describes the estimated power of the statistical test over the range of possible parameter values. The power of a statistical test is the probability of rejecting the null hypothesis when the null hypothesis is false. The estimated power is computed for all parameter values under the alternative hypothesis to create a power curve. A power analysis helps the analyst evaluate the adequacy of the sampling design when the true parameter value lies in the vicinity of the action level (which may not have been the outcome of the current study). In this manner, the analyst may determine how well a statistical test performed and compare this performance with that of other tests.

The calculations required to perform a power analysis can be relatively complicated, depending on the complexity of the sampling design and statistical test selected. Box 5.1.2 illustrates power calculations for a test of a single proportion, which is one of the simpler cases. A further discussion of power curves (performance curves) is contained in the Guidance for Data Quality Objectives (EPA QA/G-4, 1994).

Box 5.1-1: Checking Adequacy of Sample Size for a One-Sample t-Test for Simple Random Sampling

In Box 3.3-2, the one-sample t-test was used to test the hypothesis H_0: $\mu \leq 95$ ppm vs. H_A: $\mu > 95$ ppm. DQOs specified that the test should limit the false positive error rate to 5% and the false negative error rate to 20% if the true mean were 105 ppm. A random sample of size n = 9 had sample mean $\bar{x}$ = 99.38 ppm and standard deviation s = 10.41 ppm. The null hypothesis was not rejected. Assuming that the true value of the standard deviation was equal to its sample estimate 10.41 ppm, it was found that a sample size of 9 would be required, which validated the sample size of 9 which had actually been used.

The distribution of the sample standard deviation is skewed with a long right tail. It follows that the chances are greater than 50% that the sample standard deviation will underestimate the true standard deviation. In such a case it makes sense to build in some conservatism, for example, by using an upper 90% confidence limit for σ in step 5 of Box 3.3-1. Using Boxes 4.6-1 and 4.6-2 and n - 1 = 8 degrees of freedom, it is found that U = 3.49, so that an upper 90% confidence limit for the true standard deviation is

$$s\sqrt{[(n-1)/U]} = 10.41\sqrt{8/3.49} = 15.76$$

Using this value for s in Step 5 of Box 3.3-1 or Box 3.3-2 leads to the sample size estimate of 17. Hence, a sample size of at least 17 should be used to be 90% sure of achieving the DQOs. Since it is generally desirable to avoid the need for additional sampling, it is advisable to conservatively estimate sample size in the first place. In cases where DQOs depend on a variance estimate, this conservatism is achieved by intentionally overestimating the variance.

Box 5.1-2: Example of Power Calculations for the One-Sample Test of a Single Proportion

This box illustrates power calculations for the test of H_0: $P \geq .20$ vs. H_A: $P < .20$, with a false positive error rate of 5% when P=.20 presented in Boxes 3.3-9 and 3.3-10. The power of the test will be calculated assuming $P_1 = .15$ and before any data are available. Since nP_1 and $n(1-P_1)$ both exceed 4, the sample size is large enough for the normal approximation, and the test can be carried out as in steps 3 and 4 of Box 3.3-9.

STEP 1: Determine the general conditions for rejection of the null hypothesis. In this case, the null hypothesis is rejected if the sample proportion is sufficiently smaller than P_0. (Clearly, a sample proportion above P_0 cannot cast doubt on H_0.) By steps 3 and 4 of Box 3.3-9 and 3.3-10, H_0 is rejected if

$$\frac{p + .5/n - P_0}{\sqrt{P_0 Q_0 / n}} < -z_{1-\alpha}.$$

Here p is the sample proportion, $Q_0 = 1 - P_0$, n is the sample size, and $z_{1-\alpha}$ is the critical value such that $100(1-\alpha)\%$ of the standard normal distribution is below $z_{1-\alpha}$. This inequality is true if

$$p + .5/n < P_0 - z_{1-\alpha}\sqrt{P_0 Q_0 / n}.$$

STEP 2: Determine the specific conditions for rejection of the null hypothesis if P_1 (=1-Q_1) is the true value of the proportion P. The same operations as are used in step 3 of Box 3.3-9 are performed on both sides of the above inequality. However, P_0 is replaced by P_1 since it is assumed that P_1 is the true proportion. These operations make the normal approximation applicable. Hence, rejection occurs if

$$\frac{p + .5/n - P_1}{\sqrt{P_1 Q_1 / n}} < \frac{P_0 - P_1 - z_{1-\alpha}\sqrt{P_0 Q_0 / n}}{\sqrt{P_1 Q_1 / n}} = \frac{.20 - .15 - 1.645\sqrt{(.2)(.8)/85}}{\sqrt{(.15)(.85)/85}} = -0.55$$

STEP 3: Find the probability of rejection if P_1 is the true proportion. By the same reasoning that led to the test in steps 3 and 4 of Boxes 3.3-9 and 3.3-10, the quantity on the left-hand side of the above inequality is a standard normal variable. Hence the power at $P_1 = .15$ (i.e., the probability of rejection of H_0 when .15 is the true proportion) is the probability that a standard normal variable is less than -0.55. In this case, the probability is approximately 0.3 (using the last line from Table 1 of Appendix A) which is fairly small.

5.2 INTERPRETING AND COMMUNICATING THE TEST RESULTS

Sometimes difficulties may arise in interpreting or explaining the results of a statistical test. One reason for such difficulties may stem from inconsistencies in terminology; another may be due to a lack of understanding of some of the basic notions underlying hypothesis tests. As an example, in explaining the results to a data user, an analyst may use different terminology than that appearing in this guidance. For instance, rather than saying that the null hypothesis was or was not rejected, analysts may report the result of a test by saying that their computer output shows a p-value of 0.12. What does this mean? Similar problems of interpretation may occur when the data user attempts to understand the practical significance of the test results or to explain the test results to others. The following paragraphs touch on some of the philosophical issues related to hypothesis testing which may help in understanding and communicating the test results.

5.2.1 Interpretation of p-Values

The classical approach for performing hypothesis tests is to prespecify the significance level of the test, i.e., the Type I decision error rate α. This rate is used to define the decision rule associated with the hypothesis test. For instance, in testing whether the population mean μ exceeds a threshold level (e.g., 100 ppm), the test statistic may depend on $\bar{x}$, an estimate of μ. Obtaining an estimate $\bar{x}$ that is greater than 100 ppm may occur simply by chance even if the true mean μ is less than or equal to 100; however, if $\bar{x}$ is "much larger" than 100 ppm, then there is only a small chance that the null hypothesis H_0 ($\mu \leq 100$ ppm) is true. Hence the decision rule might take the form "reject H_0 if $\bar{x}$ exceeds 100 + C", where C is a positive quantity that depends on α (and on the variability of $\bar{x}$). If this condition is met, then the result of the statistical test is reported as "reject H_0"; otherwise, the result is reported as "do not reject H_0." (See Box 3.3-2 for an example of a t-test.)

An alternative way of reporting the result of a statistical test is to report its p-value, which is defined as the probability, assuming the null hypothesis to be true, of observing a test result at least as extreme as that found in the sample. Many statistical software packages report p-values, rather than adopting the classical approach of using a prespecified Type I error rate. In the above example, for instance, the p-value would be the probability of observing a sample mean as large as $\bar{x}$ (or larger) if in fact the true mean was equal to 100 ppm. Obviously, in making a decision based on the p-value, one should reject H_0 when p is small and not reject it if p is large. Thus the relationship between p-values and the classical hypothesis testing approach is that one rejects H_0 if the p-value associated with the test result is less than α. If the data user had chosen the Type I error rate as 0.05 *a priori* and the analyst reported a p-value of 0.12, then the data user would report the result as "do not reject the null hypothesis;" if the p-value had been reported as 0.03, then that person would report the result as "reject the null hypothesis." An advantage of reporting p-values is that they provide a measure of the strength of evidence for or against the null hypothesis, which allows data users to establish their own Type I error rates. The significance level can be interpreted as that p-value (α) that divides "do not reject H_O" from "reject H_O."

5.2.2 "Accepting" vs. "Failing to Reject" the Null Hypothesis

As noted in the paragraphs above, the classical approach to hypothesis testing results in one of two conclusions: "reject H_0" (called a significant result) or "do not reject H_0" (a nonsignificant result). In the latter case one might be tempted to equate "do not reject H_0" with "accept H_0." This terminology is not recommended, however, because of the philosophy underlying the classical testing procedure. This philosophy places the burden of proof on the alternative hypothesis, that is, the null hypothesis is rejected only if the evidence furnished by the data convinces us that the alternative hypothesis is the more likely state

of nature. If a nonsignificant result is obtained, it provides evidence that the null hypothesis *could* sufficiently account for the observed data, but it does not imply that the hypothesis is the only hypothesis that could be supported by the data. In other words, a highly nonsignificant result (e.g., a p-value of 0.80) may indicate that the null hypothesis provides a reasonable model for explaining the data, but it does not necessarily imply that the null hypothesis is true. It may, for example, simply indicate that the sample size was not large enough to establish convincingly that the alternative hypothesis was more likely. When the phrase "accept H_O" is encountered, it must be considered as "accepted with the preceding caveats."

5.2.3 Statistical Significance vs. Practical Significance

There is an important distinction between these two concepts. Statistical significance simply refers to the result of the hypothesis test: Was the null hypothesis rejected? The likelihood of achieving a statistically significant result depends on the true value of the population parameter being tested (for example, μ), how much that value deviates from the value hypothesized under the null hypothesis (for example, μ_0), and on the sample size. This dependence on $(\mu - \mu_0)$ is depicted by the power curve associated with the test (section 5.1.3). A steep power curve can be achieved by using a large sample size; this means that there will be a high likelihood of detecting even a small difference. On the other hand, if small sample sizes are used, the power curve will be less steep, meaning that only a very large difference between μ and μ_0 will be detectable with high probability. Hence, suppose one obtains a statistically significant result but has no knowledge of the power of the test. Then it is possible, in the case of the steep power curve, that one may be declaring significance (claiming $\mu > \mu_0$, for example) when the actual difference, from a practical standpoint, may be inconsequential. Or, in the case of the slowly increasing power curve, one may not find a significant result even though a "large" difference between μ and μ_0 exists. Neither of these situations is desirable: in the former case, there has been an excess of resources expended, whereas in the latter case, a Type II error is likely and has occurred.

But how large a difference between the parameter and the null value is of real importance? This relates to the concept of practical significance. Ideally, this question is asked and answered as part of the DQO process during the planning phase of the study. Knowing the magnitude of the difference that is regarded as being of practical significance is important during the design stage because this allows one, to the extent that prior information permits, to determine a sampling plan of type and size that will make the magnitude of that difference commensurate with a difference that can be detected with high probability. From a purely statistical design perspective, this can be considered to be main purpose of the DQO process. With such planning, the likelihood of encountering either of the undesirable situations mentioned in the prior paragraph can be reduced. Box 5.2-1 contains an example of a statistically significant but fairly inconsequential difference.

5.2.4 Impact of Bias on Test Results

Bias is defined as the difference between the expected value of a statistic and a population parameter. It is relevant when the statistic of interest (e.g., a sample average $\bar{x}$) is to be used as an estimate of the parameter (e.g., the population mean μ). For example, the population parameter of interest may be the average concentration of dioxin within the given bounds of a hazardous waste site, and the statistic might be the sample average as obtained from a random sample of points within those bounds. The expected value of a statistic can be interpreted as supposing one repeatedly implemented the particular sampling design a very large number of times and calculated the statistic of interest in each case. The average of the statistic's

Box 5.2-1: Example of a Comparison of Two Variances which is Statistically but not Practically Significant

The quality control (QC) program associated with a measurement system provides important information on performance and also yields data which should be taken into account in some statistical analyses. The QC program should include QC check samples, i.e., samples of known composition and concentration which are run at regular frequencies. The term precision refers to the consistency of a measurement method in repeated applications under fixed conditions. Precision is usually equated with a standard deviation. For many purposes, the appropriate standard deviation is one which results from applying the system to the same sample over a long period of time.

This example concerns two methods for measuring ozone in ambient air, an approved method and a new candidate method. Both methods are used once per week on a weekly basis for three months. Based on 13 analyses with each method of the mid-range QC check sample at 100 ppb, the null hypothesis of the equality of the two variances will be tested with a false positive error rate of 5% or less. (If the variances are equal, then the standard deviations are equal.) Method 1 had a sample mean of 80 ppb and a standard deviation of 4 ppb. Method 2 had a mean of 90 ppb and a standard deviation of 8 ppb. The Shapiro-Wilks test did not reject the assumption of normality for either method. Applying the F-test of Box 4.5-2, the F ratio is $8^2/4^2 = 2$. Using 12 degrees of freedom for both the numerator and denominator, the F ratio must exceed 3.28 in order to reject the hypothesis of equal variances (Table A-9 of Appendix A). Since 4 > 3.28, the hypothesis of equal variances is rejected, and it is concluded that method 1 is significantly more precise than method 2.

In an industrialized urban environment, the true ozone levels at a fixed location and time of day are known to vary over a period of months with a coefficient of variation of at least 100%. This means that the ratio of the standard deviation (SD) to the mean at a given location is at least 1. For a mean of 100 ppb, the standard deviation over time for true ozone values at the location would be at least 100 ppb. Relative to this degree of variability, a difference between measurement error standard deviations of 4 or 8 ppb is negligible. The overall variance, incorporating the true process variability and measurement error, is obtained by adding the individual variances. For instance, if measurement error standard deviation is 8 ppb, then the total variance is (100 ppb)(100 ppb) + (8 ppb)(8 ppb). Taking the square root of the variance gives a corresponding total standard deviation of 100.32 ppb. For a measurement error standard deviation of 4 ppb, the total standard deviation would be 100.08 ppb. From a practical standpoint, the difference in precision between the two methods is insignificant for the given application, despite the finding that there is a statistically significant difference between the variances of the two methods.

values would then be regarded as its expected value. Let E denote the expected value of $\bar{x}$ and denote the relationship between the expected value and the parameter, μ, as $E = \mu + b$ where b is the bias. For instance, if the bias occurred due to incomplete recovery of an analyte (and no adjustment is made), then $b = (R-100)\mu/100$, where R denotes the percent recovery. Bias may also occur for other reasons, such as lack of coverage of the entire target population (e.g., if only the drums within a storage site that are easily accessible are eligible for inclusion in the sample, then inferences to the entire group of drums may be biased). Moreover, in cases of incomplete coverage, the magnitude and direction of the bias may be unknown. An example involving comparison of the biases of two measurement methods is contained in Box 5.2-2.

In the context of hypothesis testing, the impact of bias can be quite severe in some circumstances. This can be illustrated by comparing the power curve of a test when bias is not present with a power curve for the same test when bias is present. The basic influence of bias is to shift the former “no bias” curve to the right or left, depending on the direction of the bias. If the bias is constant, then the second curve will be an exact translation of the former curve; if not, there will be a change in the shape of the second curve in addition to the translation. If the existence of the bias is unknown, then the former power curve will be regarded as the curve that determines the properties of the test when in fact the second curve will be the one that actually represents the test's power. For example, in Figure 5.2-1 when the true value of the parameter is 120, the “no bias” power is 0.72 but the true power (the biased power) is only 0.4, a substantial difference.

Box 5.2-2: Example of a Comparison of Two Biases

This example is a continuation of the ozone measurement comparison described in Box 5.2-1. Let $\bar{X}$ and s_X denote the sample mean and standard deviation of measurement method 1 applied to the QC check sample, and let $\bar{Y}$ and s_Y denote the sample mean and standard deviation of method 2. Then $\bar{X}$ = 80 ppb, s_X = 4 ppb, $\bar{Y}$ = 90 ppb and s_Y = 8 ppb. The estimated biases are $\bar{X} - \tau$ = 80 - 100 = -20 ppb for method 1, and $\bar{Y} - \tau$ = 90-100 = 10 ppb for method 2, since 100 ppb is the true value τ. That is, method 1 seems to underestimate by 20 ppb, and method 2 seems to underestimate by 10 ppb. Let μ_1 and μ_2 be the underlying mean concentrations for measurement methods 1 and 2 applied to the QC check sample. These means correspond to the average results which would obtain by applying each method a large number of times to the QC check sample, over a long period of time.

A two-sample t-test (Boxes 3.3-1 and 3.3-3) can be used to test for a significant difference between these two biases. In this case, a two-tailed test of the null hypothesis H_0: $\mu_1 - \mu_2 = 0$ against the alternative H_A: $\mu_1 - \mu_2 \neq 0$ is appropriate, because there is no *a priori* reason (in advance of data collection) to suspect that one measurement method is superior to the other. (In general, hypotheses should not be tailored to data.) Note that the difference between the two biases is the same as the difference ($\mu_1 - \mu_2$) between the two underlying means of the measurement methods. The test will be done to limit the false positive error rate to 5% if the two means are equal.

STEP 1: $\bar{X}$ = 80 ppb, s_X = 4 ppb, $\bar{Y}$ = 90 ppb, s_Y = 8 ppb.

STEP 2: From Box 5.2-1, it is known that the methods have significantly different variances, so that Sattherthwaite's t-test should be used. Therefore,

$$s_{\mathrm{NE}} = \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}} = \sqrt{\frac{4^2}{13} + \frac{8^2}{13}} = 2.48$$

STEP 3:

$$f = \frac{\left[\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right]^2}{\left[\frac{s_X^4}{m^2(m-1)} + \frac{s_Y^4}{n^2(n-1)}\right]} = \frac{\left[\frac{4^2}{13} + \frac{8^2}{13}\right]^2}{\left[\frac{4^4}{13^2\ 12} + \frac{8^4}{13^2\ 12}\right]} = 17.65.$$

Rounding down to the nearest integer gives f = 17. For a two-tailed test, the critical value is $t_{1-\alpha/2} = t_{.975}$ = 2.110, from Table A-1 of Appendix A.

STEP 4:

$$t = \frac{\bar{X} - \bar{Y}}{S_{NE}} = \frac{80 - 90}{2.48} = -4.032$$

STEP 5: For a two-tailed test, compare $|t|$ with $t_{1-\alpha/2}$ = 2.11. Since 4.032 > 2.11, reject the null hypothesis and conclude that there is a significant difference between the two method biases, in favor of method 2.

This box illustrates a situation involving two measurement methods where one method is more precise, but also more biased, than the other. If no adjustment for bias is made, then for many purposes, the less biased, more variable method is preferable. However, proper bias adjustment can make both methods unbiased, so that the more precise method becomes the preferred method. Such adjustments can be based on QC check sample results, if the QC check samples are regarded as representative of environmental samples involving sufficiently similar analytes and matrices.

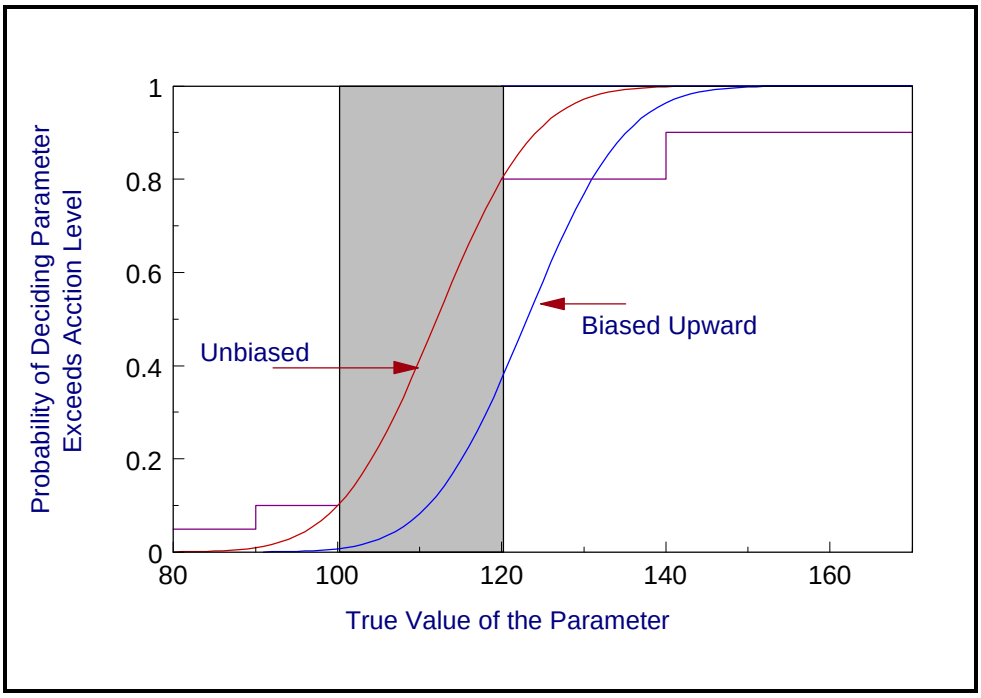


Figure 5.2-1. Illustration of Unbiased versus Biased Power Curves

Since bias is not impacted by changing the sample size, while the precision of estimates and the power of tests increases with sample size, the relative importance of bias becomes more pronounced when the sample size increases (i.e., when one makes the power curve steeper). Similarly, if the same magnitude of bias exists for two different sites, then the impact on testing errors will be more severe for the site having the smaller inherent variability in the characteristic of interest (i.e., when bias represents a larger portion of total variability).

To minimize the effects of bias: identify and document sources of potential bias; adopt measurement procedures (including specimen collection, handling, and analysis procedures) that minimize the potential for bias; make a concerted effort to quantify bias whenever possible; and make appropriate compensation for bias when possible.

5.2.5 Quantity vs. Quality of Data

The above conclusions imply that, *if compensation for bias cannot be made and if statistically-based decisions are to be made*, then there will be situations in which serious consideration should be given to using an imprecise (and perhaps relatively inexpensive) chemical method having negligible bias as compared to using a very precise method that has even a moderate degree of bias. The tradeoff favoring the imprecise method is especially relevant when the inherent variability in the population is very large relative to the random measurement error.

For example, suppose a mean concentration for a given spatial area (site) is of interest and that the coefficient of variation (CV) characterizing the site's variability is 100%. Let method A denote an imprecise method, with measurement-error CV of 40%, and let method B denote a highly precise method, with measurement-error CV of 5%. The overall variability, or total variability, can essentially be regarded as the sum of the spatial variability and the measurement variability. These are obtained from the individual CVs in

the form of variances. As CV equals standard deviation divided by mean, it follows that the site standard deviation is then the CV times the mean. Thus, for the site, the variance is 1.00^2 x mean^2; for method A, the variance is 0.40^2 x mean^2; and for method B, the variance is 0.05^2 x mean^2. The overall variability when using method A is then (1.00^2 x mean^2) + (0.40^2 x mean^2) = 1.16 x mean^2, and when using method B, the variance is (1.00^2 x mean^2) + (0.05^2 x mean^2) = 1.0025 mean^2. It follows that the overall CV when using each method is then (1.077 x mean) / mean = 107.7% for method A, and (1.001 x mean) / mean = 100.1% for method B.

Now consider a sample of 25 specimens from the site. The *precision* of the sample mean can then be characterized by the relative standard error (RSE) of the mean (which for the simple random sample situation is simply the overall CV divided by the square root of the sample size). For Method A, RSE = 21.54%; for method B, RSE = 20.02%. Now suppose that the imprecise method (Method A) is unbiased, while the precise method (Method B) has a 10% bias (e.g., an analyte percent recovery of 90%). An overall measure of error that reflects how well the sample mean estimates the site mean is the relative root mean squared error (RRMSE):

$$RRMSE = \sqrt{(RB)^2 + (RSE)^2}$$

where RB denotes the relative bias (RB = 0 for Method A since it is unbiased and RB = ±10% for Method B since it is biased) and RSE is as defined above. The overall error in the estimation of the population mean (the RRMSE) would then be 21.54% for Method A and 22.38% for Method B. If the relative bias for Method B was 15% rather than 10%, then the RRMSE for Method A would be 21.54% and the RRMSE for Method B would be 25.02%, so the method difference is even more pronounced. While the above illustration is portrayed in terms of estimation of a mean based on a simple random sample, the basic concepts apply more generally.

This example serves to illustrate that a method that may be considered preferable from a chemical point of view (e.g., 85 or 90% recovery, 5% relative standard deviation [RSD]) may not perform as well in a statistical application as a method with less bias and greater imprecision (e.g., zero bias, 40% RSD), especially when the inherent site variability is large relative to the measurement-error RSD.

5.2.6 "Proof of Safety" vs. "Proof of Hazard"

Because of the basic hypothesis testing philosophy, the null hypothesis is generally specified in terms of the *status quo* (e.g., no change or action will take place if null hypothesis is not rejected). Also, since the classical approach exercises direct control over the Type I error rate, this rate is generally associated with the error of most concern (for further discussion of this point, see section 1.2). One difficulty, therefore, may be obtaining a consensus on which error should be of most concern. It is not unlikely that the Agency's viewpoint in this regard will differ from the viewpoint of the regulated party. In using this philosophy, the Agency's ideal approach is not only to set up the direction of the hypothesis in such a way that controlling the Type I error protects the health and environment but also to set it up in a way that encourages quality (high precision and accuracy) and minimizes expenditure of resources in situations where decisions are relatively "easy" (e.g., all observations are far from the threshold level of interest).

In some cases, how one formulates the hypothesis testing problem can lead to very different sampling requirements. For instance, following remediation activities at a hazardous waste site, one may seek to answer "Is the site clean?" Suppose one attempts to address this question by comparing a mean level from samples taken after the remediation with a threshold level (chosen to reflect "safety"). If the threshold level is

near background levels that might have existed in the absence of the contamination, then it may be very difficult (i.e., require enormous sample sizes) to "prove" that the site is "safe." This is because the concentrations resulting from even a highly efficient remediation under such circumstances would not be expected to deviate greatly from such a threshold. A better approach for dealing with this problem may be to compare the remediated site with a reference ("uncontaminated") site, assuming that such a site can be determined.

To avoid excessive expense in collecting and analyzing samples for a contaminant, compromises will sometimes be necessary. For instance, suppose that a significance level of 0.05 is to be used; however, the affordable sample size may be expected to yield a test with power of only 0.40 at some specified parameter value chosen to have practical significance (see section 5.2.3). One possible way that compromise may be made in such a situation is to relax the significance level, for instance, using $\alpha = 0.10$, 0.15, or 0.20. By relaxing this false positive rate, a higher power (i.e., a lower false negative rate β) can be achieved. An argument can be made, for example, that one should develop sampling plans and determine sample sizes in such a way that both the Type I and Type II errors are treated simultaneously and in a balanced manner (for example, designing to achieve $\alpha = \beta = 0.15$) instead of using the traditional approach of fixing the Type I error rate at 0.05 or 0.01 and letting β be determined by the sample size. This approach of treating the Type I and Type II errors simultaneously is taken in the DQO Process and it is recommended that several different scenarios of α and β be investigated before a decision on specific values for α and β are selected.

APPENDIX A

STATISTICAL TABLES

LIST OF TABLES

Table No. **Page**

TABLE A-1: CRITICAL VALUES OF STUDENT'S t DISTRIBUTION . . . . . A - 3

TABLE A-2: CRITICAL VALUES FOR THE STUDENTIZED RANGE TEST . . . . . A - 4

TABLE A-3: CRITICAL VALUES FOR THE EXTREME VALUE TEST (DIXON'S TEST) . . . . . A - 5

TABLE A-4: CRITICAL VALUES FOR DISCORDANCE TEST . . . . . A - 6

TABLE A-5: APPROXIMATE CRITICAL VALUES λ_r FOR ROSNER'S TEST . . . . . A - 7

TABLE A-6: QUANTILES OF THE WILCOXON SIGNED RANKS TEST . . . . . A - 9

TABLE A-7: CRITICAL VALUES FOR THE RANK-SUM TEST - $\alpha = 0.05$ . . . . . A - 10

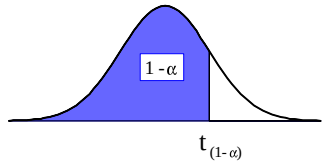
TABLE A-8: PERCENTILES OF THE CHI-SQUARE DISTRIBUTION . . . . . A - 11

TABLE A-9: PERCENTILES OF THE F DISTRIBUTION . . . . . A - 12

TABLE A-10: VALUES OF THE PARAMETER $\hat{\lambda}$ FOR COHEN'S ESTIMATES . . . . . A - 15

TABLE A-11: PROBABILITIES FOR THE SMALL-SAMPLE MANN-KENDALL TEST FOR TREND . . . . . A - 16

TABLE A-1: CRITICAL VALUES OF STUDENT'S t DISTRIBUTION



Degrees of Freedom	1 - α								
	.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

Note: The last row of the table (∞ degrees of freedom) gives the criticial values for a standard normal distribution (z), e.g., $t_{\infty, 0.95} = z_{0.95} = 1.645$.

TABLE A-2: CRITICAL VALUES FOR THE STUDENTIZED RANGE TEST

	Level of Significance α					
	0.01		**0.05**		**0.10**	
n	*a*	*b*	*a*	*b*	*a*	*b*
3	1.737	2.000	1.758	1.999	1.782	1.997
4	1.87	2.445	1.98	2.429	2.04	2.409
5	2.02	2.803	2.15	2.753	2.22	2.712
6	2.15	3.095	2.28	3.012	2.37	2.949
7	2.26	3.338	2.40	3.222	2.49	3.143
8	2.35	3.543	2.50	3.399	2.59	3.308
9	2.44	3.720	2.59	3.552	2.68	3.449
10	2.51	3.875	2.67	3.685	2.76	3.57
11	2.58	4.012	2.74	3.80	2.84	3.68
12	2.64	4.134	2.80	3.91	2.90	3.78
13	2.70	4.244	2.86	4.00	2.96	3.87
14	2.75	4.34	2.92	4.09	3.02	3.95
15	2.80	4.44	2.97	4.17	3.07	4.02
16	2.84	4.52	3.01	4.24	3.12	4.09
17	2.88	4.60	3.06	4.31	3.17	4.15
18	2.92	4.67	3.10	4.37	3.21	4.21
19	2.96	4.74	3.14	4.43	3.25	4.27
20	2.99	4.80	3.18	4.49	3.29	4.32
25	3.15	5.06	3.34	4.71	3.45	4.53
30	3.27	5.26	3.47	4.89	3.59	4.70
35	3.38	5.42	3.58	5.04	3.70	4.84
40	3.47	5.56	3.67	5.16	3.79	4.96
45	3.55	5.67	3.75	5.26	3.88	5.06
50	3.62	5.77	3.83	5.35	3.95	5.14
55	3.69	5.86	3.90	5.43	4.02	5.22
60	3.75	5.94	3.96	5.51	4.08	5.29
65	3.80	6.01	4.01	5.57	4.14	5.35
70	3.85	6.07	4.06	5.63	4.19	5.41
75	3.90	6.13	4.11	5.68	4.24	5.46
80	3.94	6.18	4.16	5.73	4.28	5.51
85	3.99	6.23	4.20	5.78	4.33	5.56
90	4.02	6.27	4.24	5.82	4.36	5.60
95	4.06	6.32	4.27	5.86	4.40	5.64
100	4.10	6.36	4.31	5.90	4.44	5.68
150	4.38	6.64	4.59	6.18	4.72	5.96
200	4.59	6.84	4.78	6.39	4.90	6.15
500	5.13	7.42	5.47	6.94	5.49	6.72
1000	5.57	7.80	5.79	7.33	5.92	7.11

TABLE A-3: CRITICAL VALUES FOR THE EXTREME VALUE TEST (DIXON'S TEST)

	Level of Significance α		
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-4: CRITICAL VALUES FOR DISCORDANCE TEST

	Level of Significance α	
n	**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

	Level of Significance α	
n	**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-5: APPROXIMATE CRITICAL VALUES λ_r FOR ROSNER'S TEST

		α	
n	**r**	**0.05**	**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

		α	
n	**r**	**0.05**	**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

		α	
n	**r**	**0.05**	**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

TABLE A-5: APPROXIMATE CRITICAL VALUES λ_r FOR ROSNER'S TEST

		α	
n	**r**	**0.05**	**0.01**
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
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

		α	
n	**r**	**0.05**	**0.01**
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
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

		α	
n	**r**	**0.05**	**0.01**
250	1	3.67	4.04
	5	3.67	4.04
	10	3.66	4.03
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-6: QUANTILES OF THE WILCOXON SIGNED RANKS TEST

n	$w_{.01}$	$w_{.05}$	$w_{.10}$	$w_{.20}$
4	0	0	1	3
5	0	1	3	4
6	0	3	4	6
7	1	4	6	9
8	2	6	9	12
9	4	9	11	15
10	6	11	15	19
11	8	14	18	23
12	10	18	22	28
13	13	22	27	33
14	16	26	32	39
15	20	31	37	45
16	24	36	43	51
17	28	42	49	58
18	33	48	56	66
19	38	54	63	74
20	44	61	70	82

TABLE A-7: CRITICAL VALUES FOR THE RANK-SUM TEST - $\alpha = 0.05$

Smaller of *m* or *n*	Larger of *m* or *n*																	
	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																	0	0
2			0	0	0	1	1	1	1	2	2	3	3	3	3	4	4	4
3	0	0	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	11
4		1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5			4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6				7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7					11	13	15	17	19	21	24	26	28	30	33	35	37	39
8						15	18	20	23	26	28	31	33	36	39	41	44	47
9							21	24	27	30	33	36	39	42	45	48	51	54
10								27	31	34	37	41	44	48	51	55	58	62
11									34	38	42	46	50	54	57	61	65	69
12										42	47	51	55	60	64	68	72	77
13											51	56	61	65	70	75	80	84
14												61	66	71	77	82	87	92
15													72	77	83	88	94	100
16														83	89	95	101	107
17															96	102	109	115
18																109	116	123
19																	123	130
20																		138

TABLE A-8: PERCENTILES OF THE CHI-SQUARE DISTRIBUTION

v	1 - α									
	.005	.010	.025	.050	.100	.900	.950	.975	.990	.995
1	$0.0^{4}393$	$0.0^{3}157$	$0.0^{3}982$	$0.0^{2}393$	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	3.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	53.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

TABLE A-9: PERCENTILES OF THE F DISTRIBUTION

Degrees Freedom for Denom-inator		Degrees of Freedom for Numerator																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	60	120	∞
1	.50	1.00	1.50	1.71	1.82	1.89	1.94	1.98	2.00	2.03	2.04	2.07	2.09	2.12	2.13	2.15	2.17	2.18	2.20
	.90	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.7	61.2	61.7	62.0	62.3	62.8	63.1	63.3
	.95	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	252	253	254
	.975	648	800	864	900	922	937	948	957	963	969	977	985	993	997	1001	1010	1014	1018
	.99	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6313	6339	6366
2	.50	0.667	1.00	1.13	1.21	1.25	1.28	1.30	1.32	1.33	1.34	1.36	1.38	1.39	1.40	1.41	1.43	1.43	1.44
	.90	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.48	9.49
	.95	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
	.975	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5
	.99	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5
3	.50	0.585	0.881	1.00	1.06	1.10	1.13	1.15	1.16	1.17	1.18	1.20	1.21	1.23	1.23	1.24	1.25	1.26	1.27
	.90	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.15	5.14	5.13
	.95	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.57	8.55	8.53
	.975	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3	14.2	14.1	14.1	14.0	13.9	13.9
	.99	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.3	26.2	26.1
4	.50	0.549	0.828	0.941	1.00	1.04	1.06	1.08	1.09	1.10	1.11	1.13	1.14	1.15	1.16	1.16	1.18	1.18	1.19
	.90	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.79	3.78	3.76
	.95	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.69	5.66	5.63
	.975	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.36	8.31	8.26
	.99	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.6	13.5
	.999	74.1	61.2	56.2	53.4	51.7	50.5	49.7	49.0	48.5	48.1	47.4	46.8	46.1	45.8	45.4	44.7	44.4	44.1
5	.50	0.528	0.799	0.907	0.965	1.00	1.02	1.04	1.05	1.06	1.07	1.09	1.10	1.11	1.12	1.12	1.14	1.14	1.15
	.90	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.39	3.27	3.24	3.21	3.19	3.17	3.14	3.12	3.11
	.95	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.43	4.40	4.37
	.975	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.12	6.07	6.02
	.99	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.20	9.11	9.02
	.999	47.2	37.1	33.2	31.1	29.8	28.8	28.2	27.6	27.2	26.9	26.4	25.9	25.4	25.1	24.9	24.3	24.1	23.8
6	.50	0.515	0.780	0.886	0.942	0.977	1.00	1.02	1.03	1.04	1.05	1.06	1.07	1.08	1.09	1.10	1.11	1.12	1.12
	.90	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.76	2.74	2.72
	.95	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.74	3.70	3.67
	.975	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	4.96	4.90	4.85
	.99	22.8	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.06	6.97	6.88
	.999	35.5	27.0	23.7	21.9	20.8	20.0	19.5	19.0	18.7	18.4	18.0	17.6	17.1	16.9	16.7	16.2	16.0	15.7

TABLE A-9: PERCENTILES OF THE F DISTRIBUTION

Degrees Freedom for Denom-inator		Degrees of Freedom for Numerator																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	60	120	∞
7	.50	0.506	.0767	0.871	0.926	0.960	0.983	1.00	1.01	1.02	1.03	1.04	1.05	1.07	1.07	1.08	1.09	1.10	1.10
	.90	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.51	2.49	2.47
	.95	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.30	3.27	3.23
	.975	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.25	4.20	4.14
	.99	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.82	5.74	5.65
	.999	29.2	21.7	18.8	17.2	16.2	15.5	15.0	14.6	14.5	14.1	13.7	13.3	12.9	12.7	12.5	12.1	11.9	11.7
8	.50	0.499	0.757	0.860	0.915	0.948	0.971	0.988	1.00	1.01	1.02	1.03	1.04	1.05	1.06	1.07	1.08	1.08	1.09
	.90	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.34	2.32	2.29
	.95	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.01	2.97	2.93
	.975	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.78	3.73	3.67
	.99	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.03	4.95	4.86
	.999	25.4	18.5	15.8	14.4	13.5	12.9	12.4	12.0	11.8	11.5	11.2	10.8	10.5	10.3	10.1	9.73	9.53	9.33
9	.50	0.494	0.749	0.852	0.906	0.939	0.962	0.978	0.990	1.00	1.01	1.01	1.03	1.04	1.05	1.05	1.07	1.07	1.08
	.90	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.21	2.18	2.16
	.95	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.79	2.75	2.71
	.975	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.45	3.39	3.33
	.99	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.48	4.40	4.31
	.999	22.9	16.4	13.9	12.6	11.7	11.1	10.7	10.4	10.1	9.89	9.57	9.24	8.90	8.72	8.55	8.19	8.00	7.81
10	.50	0.490	0.743	0.845	0.899	0.932	0.954	0.971	0.983	0.992	1.00	1.01	1.02	1.03	1.04	1.05	1.06	1.06	1.07
	.90	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.11	2.08	2.06
	.95	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.74	2.70	2.62	2.58	2.54
	.975	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.20	3.14	3.08
	.99	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.08	4.00	3.91
	.999	21.0	14.9	12.6	11.3	10.5	9.93	9.52	9.20	8.96	8.75	8.45	8.13	7.80	7.64	7.47	7.12	6.94	6.76
12	.50	0.484	0.735	0.835	0.888	0.921	0.943	0.959	0.972	0.981	0.989	1.00	1.01	1.02	1.03	1.03	1.05	1.05	1.06
	.90	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.96	1.93	1.90
	.95	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.38	2.34	2.30
	.975	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.85	2.79	2.72
	.99	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.54	3.45	3.36
	.999	18.6	13.0	10.8	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.71	6.40	6.25	6.09	5.76	5.59	5.42
15	.50	0.478	0.726	0.826	0.878	0.911	0.933	0.949	0.960	0.970	0.977	0.989	1.00	1.01	1.02	1.02	1.03	1.04	1.05
	.90	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.82	1.79	1.76
	.95	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.16	2.11	2.07
	.975	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.52	2.46	2.40
	.99	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.05	2.96	2.87
	.999	16.6	11.3	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.81	5.54	5.25	5.10	4.95	4.64	4.48	4.31

TABLE A-9: PERCENTILES OF THE F DISTRIBUTION

Degrees Freedom for Denom-inator		Degrees of Freedom for Numerator																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	60	120	∞
20	.50	0.472	0.718	0.816	0.868	0.900	0.922	0.938	0.950	0.959	0.966	0.977	0.989	1.00	1.01	1.01	1.02	1.03	1.03
	.90	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.68	1.64	1.61
	.95	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.95	1.90	1.84
	.975	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.22	2.16	2.09
	.99	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.61	2.52	2.42
	.999	14.8	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.82	4.56	4.29	4.15	4.00	3.70	3.54	3.38
24	.50	0.469	0.714	0.812	0.863	0.895	0.917	0.932	0.944	0.953	0.961	0.972	0.983	0.994	1.00	1.01	1.02	1.02	1.03
	.90	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.61	1.57	1.53
	.95	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.84	1.79	1.73
	.975	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.08	2.01	1.94
	.99	7.82	6.66	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.40	2.31	2.21
	.999	14.0	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64	4.39	4.14	3.87	3.74	3.59	3.29	3.14	2.97
30	.50	0.466	0.709	0.807	0.858	0.890	0.912	0.927	0.939	0.948	0.955	0.966	0.978	0.989	0.994	1.00	1.01	1.02	1.02
	.90	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.62	1.64	1.61	1.54	1.50	1.46
	.95	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.74	1.68	1.62
	.975	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	1.94	1.87	1.79
	.99	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.21	2.11	2.01
	.999	13.3	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.00	3.75	3.49	3.36	3.22	2.92	2.76	2.59
60	.50	0.461	0.701	0.798	0.849	0.880	0.901	0.917	0.928	0.937	0.945	0.956	0.967	0.978	0.983	0.989	1.00	1.01	1.01
	.90	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.40	1.35	1.29
	.95	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.53	1.47	1.39
	.975	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.67	1.58	1.48
	.99	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.84	1.73	1.60
	.999	12.0	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54	3.32	3.08	2.83	2.69	2.55	2.25	2.08	1.89
120	. 90	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.32	1.26	1.19
	.95	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.43	1.35	1.25
	.975	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.95	1.82	1.76	1.69	1.53	1.43	1.31
	.99	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.66	1.53	1.38
	.999	11.4	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24	3.02	2.78	2.53	2.40	2.26	1.95	1.77	1.54
∞	.90	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.24	1.17	1.00
	.95	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.32	1.22	1.00
	.975	5.02	3.69	3.12	2.79	2.57	22.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.39	1.27	1.00
	.99	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.47	1.32	1.00
	.999	10.8	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96	2.74	2.51	2.27	2.13	1.99	1.66	1.45	1.00

TABLE A-10: VALUES OF THE PARAMETER $\hat{\lambda}$ FOR COHEN'S ESTIMATES ADJUSTING FOR NONDETECTED VALUES

	h											
γ	.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	.27626
.30	.012243	.024658	.037249	.050018	.062969	.076106	.089433	.10295	.11667	.13059	.20338	.28193
.35	.012520	.025211	.038077	.051120	.064345	.077736	.091355	.10515	.11914	.13333	.20747	.28737
.40	.012784	.025738	.038866	.052173	.065660	.079332	.093193	.10725	.12150	.13595	.21129	.29250
.45	.013036	.026243	.039624	.053182	.066921	.080845	.094958	.10926	.12377	.13847	.21517	.29765
.50	.013279	.026728	.040352	.054153	.068135	.082301	.096657	.11121	.12595	.14090	.21882	.30253
.55	.013513	.027196	.041054	.055089	.069306	.083708	.098298	.11208	.12806	.14325	.22225	.30725
.60	.013739	.027849	.041733	.055995	.070439	.085068	.099887	.11490	.13011	.14552	.22578	.31184
.65	.013958	.028087	.042391	.056874	.071538	.086388	.10143	.11666	.13209	.14773	.22910	.31630
.70	.014171	.028513	.043030	.057726	.072505	.087670	.10292	.11837	.13402	.14987	.23234	.32065
.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	.015338	.030850	.046540	.062413	.078471	.094720	.11116	.12780	.14465	.16170	.25022	.34471

	h											
γ	.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-11: PROBABILITIES FOR THE SMALL-SAMPLE MANN-KENDALL TEST FOR TREND

	n					n		
S	4	5	8	9	**S**	6	7	10
0	0.625	0.592	0.548	0.540	1	0.500	0.500	0.500
2	0.375	0.408	0.452	0.460	3	0.360	0.386	0.431
4	0.167	0.242	0.360	0.381	5	0.235	0.281	0.364
6	0.042	0.117	0.274	0.306	7	0.136	0.191	0.300
8		0.042	0.199	0.238	9	0.068	0.199	0.242
10		0.0083	0.138	0.179	11	0.028	0.068	0.190
12			0.089	0.130	13	0.0083	0.035	0.146
14			0.054	0.090	15	0.0014	0.015	0.108
16			0.031	0.060	17		0.0054	0.078
18			0.016	0.038	19		0.0014	0.054
20			0.0071	0.022	21		0.00020	0.036
22			0.0028	0.012	23			0.023
24			0.00087	0.0063	25			0.014
26			0.00019	0.0029	27			0.0083
28			0.000025	0.0012	29			0.0046
30				0.00043	31			0.0023
32				0.00012	33			0.0011
34				0.000025	35			0.00047
36				0.0000028	37			0.00018
					39			0.0000458
					41			0.0000415
					43			0.0000028
					45			0.00000028

APPENDIX B

REFERENCES

APPENDIX B: REFERENCES

		Page
B.1	CHAPTER 1	B - 3
B.2	CHAPTER 2	B - 4
B.3	CHAPTER 3	B - 4
B.4	CHAPTER 4	B - 4
B.5	CHAPTER 5	B - 5
B.6	LIST OF REFERENCES	B - 5
	B.6.1 Primary References	B - 5
	B.6.2 Basic Statistics Textbooks	B - 5
	B.6.3 Secondary References	B - 5

LIST OF TABLES

Table No.	Page
B-1: Selected References from Primary and Introductory Textbooks	B - 8
B-2: Secondary References	B - 9

APPENDIX B: REFERENCES

This appendix provides references for the topics and procedures described in this document. The references are broken into three groups: Primary, Basic Statistics Textbooks, and Secondary. This classification does not refer in any way to the subject matter content but to the relevance to the intended audience for this document, ease in understanding statistical concepts and methodologies, and accessability to the non-statistical community. Primary references are those thought to be of particular benefit as hands-on material, where the degree of sophistication demanded by the writer seldom requires extensive training in statistics; most of these references should be on an environmental statistician's bookshelf. References to specific chapters within the primary references are provided in Table B-1 (at the end of this appendix) for each specific topic. Secondary references are original research works, theoretical discussions or expositions, or methodologies needing further development before being immediately adaptable to environmental problems. References for original research works are listed in Table B-2 (at the end of this appendix) for topics described in this guidance. Users of this document are encouraged to send recommendations on additional references to the address listed in the Foreword.

Some sections within the chapters reference materials found in most introductory statistics books. This document uses Walpole and Myers (1985), Freedman, Pisani, Purves, and Adhakari (1991), Mendenhall (1987), and Dixon and Massey (1983). Table B-1 (at the end of this appendix) lists specific chapters in these books where topics contained in this guidance may be found. This list could be extended much further by use of other basic textbooks; this is acknowledged by the simple statement that further information is available from introductory text books.

Some important books specific to the analysis of environmental data include: Gilbert (1987), an excellent all-round handbook having strength in sampling, estimation, and hot-spot detection; Gibbons (1994), a book specifically concentrating on the application of statistics to groundwater problems with emphasis on method detection limits, censored data, and the detection of outliers; and Madansky (1988), a slightly more theoretical volume with important chapters on the testing for Normality, transformations, and testing for independence. In addition, Ott (1995) describes modeling, probabilistic processes, and the Lognormal distribution of contaminants, and Berthouex and Brown (1994) provide an engineering approach to problems including estimation, experimental design and the fitting of models.

B.1 CHAPTER 1

Chapter 1 establishes the framework of qualitative and quantitative criteria against which the data that has been collected will be assessed. The most important feature of this chapter is the concept of the test of hypotheses framework which is described in any introductory textbook. A non-technical exposition of hypothesis testing is also to be found in U.S. EPA (1994a, 1994b) which provides guidance on planning for environmental data collection.

A full discussion of sampling methods with the attendant theory are to be found in Gilbert (1987) and a shorter discussion may be found in U.S. EPA (1989). Cochran (1966) and Kish (1965) also provide more advanced theoretical concepts but may require the assistance of a statistician for full comprehension. More sophisticated sampling designs such as composite sampling, adaptive sampling, and ranked set sampling, will be discussed in future Agency guidance.

B.2 CHAPTER 2

Standard statistical quantities and graphical representations are discussed in most introductory statistics books. In addition, Berthouex & Brown (1994) and Madansky (1988) both contain thorough discussions on the subject. There are also several textbooks devoted exclusively to graphical representations, including Cleveland (1993), which may contain the most applicable methods for environmental data, Tufte (1983), and Chambers, Cleveland, Kleiner and Tukey (1983).

Two EPA sources for temporal data that keep theoretical discussions to a minimum are U.S. EPA (1992a) and U.S. EPA (1992b). For a more complete discussion on temporal data, specifically time series analysis, see Box and Jenkins (1970), Wei (1990), or Ostrum (1978). These more complete references provide both theory and practice; however, the assistance of a statistician may be needed to adapt the methodologies for immediate use. Theoretical discussions of spatial data may be found in Journel and Huijbregts (1978), Cressie (1993), and Ripley (1981).

B.3 CHAPTER 3

The hypothesis tests covered in this edition of the guidance are well known and straight-forward; basic statistics texts cover these subjects. Future editions of this guidance will expand on these tests to include: tests for the mean of skewed distributions, tests for data from ranked set samples, and t-tests for winsorized or trimmed data.

Besides basic statistical text books, Berthouex & Brown (1994), Hardin and Gilbert (1993), and U.S. EPA (1989, 1994c) may be useful to the reader. In addition, there are some statistics books devoted specifically to hypothesis testing, for example, see Lehmann (1991). These books may be too theoretical for most practitioners, and their application to environmental situations may not be obvious.

The statement in this document that the sign test requires approximately 1.225 times as many observations as the Wilcoxon rank sum test to achieve a given power at a given significance level is attributable to Lehmann (1975).

B.4 CHAPTER 4

This chapter is essentially a compendium of statistical tests drawn mostly from the primary references and basic statistics textbooks. Gilbert (1987) and Madansky (1988) have an excellent collection of techniques and U.S. EPA (1992a) contains techniques specific to water problems.

For Normality (section 4.2), Madansky (1988) has an excellent discussion on tests as does Shapiro (1986). For trend testing (section 4.3), Gilbert (1987) has an excellent discussion on statistical tests and U.S. EPA (1992b) provides adjustments for trends and seasonality in the calculation of descriptive statistics.

There are several very good textbooks devoted to the treatment of outliers (section 4.4). Two authoritative texts are Barnett and Lewis (1978) and Hawkins (1980). Additional information is also to be found in Beckman and Cook (1983) and Tietjen and Moore (1972). Several useful software programs are available on the statistical market including U.S. EPA's *GEO-EASE* and *Scout*, both developed by the Environmental Monitoring Systems Laboratory, Las Vegas, Nevada and described in U.S. EPA (1991) and U.S. EPA (1993b), respectively.

Tests for dispersion (section 4.5) are described in the basic textbooks and examples are to be found in U.S. EPA (1992a). Transformation of data (section 4.6) is a sensitive topic and thorough discussions may be found in Gilbert (1987), and Dixon and Massey (1983). Equally sensitive is the analysis of data where some values are recorded as non-detected (section 4.7); Gibbons (1994) and U.S. EPA (1992a) have relevant discussions and examples.

B.5 CHAPTER 5

Chapter 5 discusses some of the philosophical issues related to hypothesis testing which may help in understanding and communicating the test results. Although there are no specific references for this chapter, many topics (e.g., the use of p-values) are discussed in introductory textbooks. Future editions of this guidance will be expanded by incorporating practical experiences from the environmental community into this chapter.

B.6 LIST OF REFERENCES

B.6.1 Primary References

Berthouex, P.M., and L.C. Brown, 1994. *Statistics for Environmental Engineers*. Lewis, Boca Raton, FL.

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

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

Madansky, A., 1988. *Prescriptions for Working Statisticians.* Springer-Verlag, New York, NY.

Ott, W.R., 1995. *Environmental Statistics and Data Analysis.* Lewis, Boca Raton, FL.

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

U.S. Environmental Protection Agency, 1994a. *Guidance for the Data Quality Objectives Process* (EPA QA/G4). EPA/600/R-96/055. Office of Research and Development.

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

U.S. Environmental Protection Agency , 1992a. *Guidance Document on the Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities.* EPA/530/R-93/003. Office of Solid Waste. (NTIS: PB89-151026)

B.6.2 Basic Statistics Textbooks

Dixon, W.J., and F.J. Massey, Jr., 1983. *Introduction to Statistical Analysis* (Fourth Edition). McGraw-Hill, New York, NY.

Freedman, D., R. Pisani, R. Purves, and A. Adhikari, 1991. *Statistics.* W.W. Norton & Co., New York, NY.

Mendenhall, W., 1987. *Introduction to Probability and Statistics* (Seventh Edition). PWS-Kent, Boston, MA.

Walpole, R., and R. Myers, 1985. *Probability and Statistics for Engineers and Scientists* (Third Ed.). MacMillan, New York, NY.

B.6.3 Secondary References

Barnett, V., and T. Lewis, 1978. *Outliers in Statistical Data.* John Wiley, New York, NY.

Beckman, R.J., and R.D. Cook, 1983. Outlier.....s, *Technometrics* **25**:119-149.

Box, G.E.P., and G.M. Jenkins, 1970. *Time Series Analysis, Forecasting, and Control.* Holden-Day, San Francisco, CA.

Chambers, J.M, W.S. Cleveland, B. Kleiner, and P.A. Tukey, 1983. *Graphical Methods for Data Analysis.* Wadsworth & Brooks/Cole Publishing Co., Pacific Grove, CA.

Cleveland, W.S., 1993. *Visualizing Data*. Hobart Press, Summit, NJ.

Cochran, W. G., 1966. *Sampling Techniques* (Third Edition). John Wiley, New York, NY.

Cohen, A.C., Jr. 1959. Simplified estimators for the normal distribution when samples are singly censored or truncated, *Technometrics* 1:217-237.

Conover, W.J., 1980. *Practical Nonparametric Statistics* (Second Edition). John Wiley, New York, NY.

Cressie, N., 1993. *Statistics for Spatial Data.* John Wiley, New York, NY.

D'Agostino, R.B., 1971. An omnibus test of normality for moderate and large size samples, *Biometrika* **58:**341-348.

David, H.A., H.O. Hartley, and E.S. Pearson, 1954. The distribution of the ratio, in a single normal sample, of range to standard deviation, *Biometrika* **48:**41-55.

Dixon, W.J., 1953. Processing data for outliers, *Biometrika* **9:**74-79.

Filliben, J.J., 1975. The probability plot correlation coefficient test for normality, *Technometrics* **17:**111-117.

Geary, R.C., 1947. Testing for normality, *Biometrika* **34:**209-242.

Geary, R.C., 1935. The ratio of the mean deviation to the standard deviation as a test of normality, *Biometrika* **27:**310-32.

Grubbs, F.E., 1969. Procedures for detecting outlying observations in samples, *Technometrics* **11:**1-21.

Hardin, J.W., and R.O. Gilbert, 1993. *Comparing Statistical Tests for Detecting Soil Contamination Greater than Background, Report to U.S. Department of Energy,* PNL-8989, UC-630, Pacific Northwest Laboratory, Richland, WA.

Hawkins, D.M., 1980. *Identification of Outliers.* Chapman and Hall, New York, NY.

Journel, A.G., and C.J. Huijbregts, 1978. *Mining Geostatistics.* Academic Press, London.

Kish, L., 1965. *Survey Sampling*. John Wiley, New York, NY.

Kleiner, B., and J.A. Hartigan, 1981. Representing points in many dimensions by trees and castles, *Journal of the American Statistical Association* **76:**260.

Lehmann, E.L., 1991. *Testing Statistical Hypotheses*. Wadsworth & Brooks/Cole Publishing Co., Pacific Grove, CA.

Lehmann, E.L., 1975. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, Inc., San Francisco, CA.

Lilliefors, H.W., 1969. Correction to the paper "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association* **64:**1702.

Lilliefors, H.W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *Journal of the American Statistical Association* **64:**399-402.

Ostrum, C.W., 1978. *Time Series Analysis* (Second Edition). Sage University Papers Series, Vol 9. Beverly Hills and London.

Ripley, B.D., 1981. *Spatial Statistics.* John Wiley and Sons, Somerset, NJ.

Rosner, B., 1975. On the detection of many outliers, *Technometrics* **17:**221-227.

Royston, J.P., 1982. An extension of Shapiro and Wilk's W test for normality to large samples, *Applied Statistics* **31:**161-165.

Sen, P.K., 1968a. Estimates of the regression coefficient based on Kendall's tau, *Journal of the American Statistical Association* **63:**1379-1389.

Sen, P.K., 1968b. On a class of aligned rank order tests in two-way layouts, *Annals of Mathematical Statistics* **39:**1115-1124.

Shapiro, S., 1986. *Volume 3: How to Test Normality and Other Distributional Assumptions.* American Society for Quality Control, Milwaukee, WI.

Shapiro, S., and M.B. Wilk, 1965. An analysis of variance test for normality (complete samples), *Biometrika* **52:**591-611.

Siegel, J.H., R.W. Goldwyn, and H.P. Friedman, 1971. Pattern and process of the evolution of human septic shock, *Surgery* **70:**232.

Stefansky, W., 1972. Rejecting outliers in factorial designs, *Technometrics* **14:**469-478.

Tietjen, G.L., and R.M. Moore, 1972. Some Grubbs-type statistics for the detection of several outliers, *Technometrics* **14:**583-597.

Tufte, E.R., 1983. *The Visual Display of Quantitative Information.* Graphics Press, Chesire, CN.

Tufte, E.R., 1990. *Envisioning Information.* Graphics Press, Chesire, CN.

U. S. Environmental Protection Agency, 1994c. *Methods for Evaluating the Attainments of Cleanup Standards: Volume 3: Reference-Based Standards*. EPA/230/R-94-004. Office of Policy, Planning, and Evaluation. (NTIS: PB94-176831)

U.S. Environmental Protection Agency, 1993a. *The Data Quality Objectives Process for Superfund: Interim Final Guidance.* EPA/540/R-93/071. Office of Emergency and Remedial Response.

U.S. Environmental Protection Agency, 1993b. *Scout: A Data Analysis Program.* Environmental Monitoring Systems Laboratory, Office of Research and Development. (NTIS: PB93-505303)

U. S. Environmental Protection Agency, 1992b. *Methods for Evaluating the Attainments of Cleanup Standards: Volume 2: Ground Water.* EPA/230/R-92/014. Office of Policy, Planning, and Evaluation. (NTIS: PB94-138815)

U.S. Environmental Protection Agency, 1991. *GEO-EAS 1.2.1. User's Guide.* EPA 600/8-91/008. Environmental Monitoring Systems Laboratory, Office of Research and Development. (NTIS: PB93-504967)

U. S. Environmental Protection Agency, 1989. *Methods for Evaluating the Attainments of Cleanup Standards: Volume 1: Soils and Solid Media.* EPA/230/02-89-042. Office of Policy, Planning, and Evaluation. (NTIS: PB89-234959*)*

Walsh, J.E., 1958. Large sample nonparametric rejection of outlying observations, *Annals of the Institute of Statistical Mathematics* **10:**223-232.

Walsh, J.E., 1953. Correction to "Some nonparametric tests of whether the largest observations of a set are too large or too small," *Annals of Mathematical Statistics* **24:**134-135.

Walsh, J.E., 1950. Some nonparametric tests of whether the largest observations of a set are too large or too small, *Annals of Mathematical Statistics* **21:**583-592.

Wang, Peter C.C., 1978. *Graphical Representation of Multivariate Data.* Academic Press, New York, NY.

Wegman, Edward J., 1990. Hyperdimensional data analysis using parallel coordinates, *Journal of the American Statistical Association* **85:** 664.

Wei, W.S., 1990. *Time Series Analysis (Second Edition)*. Addison Wesley, Menlo Park, CA.

Table B-1: Selected References from Primary and Introductory Textbooks

Subject	Section	Source (with Chapter)
Measures of Relative Standing	2.2.1	Dixon & Massey 2-3
Measures of Central Tendency	2.2.2	Walpole & Myers 6.4, Dixon & Massey 3-1
Measures of Dispersion	2.2.3	Walpole & Myers 6.4, Dixon & Massey 3-2
Measures of Association	2.2.4	Walpole & Myers 9.9, Dixon & Massey 11-6
Histogram/Frequency Plots	2.3.1	Walpole & Myers 2.4, Dixon & Massey 2-2
Stem-and-Leaf Plot	2.3.2	Walpole & Myers 2.4
Quantile Plot	2.3.5	Walpole & Myers 2.4
Normal Probability Plot (Q-Q Plot)	2.3.6	Dixon & Massey 5-4
One-Sample t-Test	3.2.1	Walpole & Myers 8.4, Dixon & Massey 7-1
Wilcoxon Signed Rank Test	3.2.1	Walpole & Myers 14.3, Dixon & Massey 17-2
One-Sample Proportion Test	3.2.2	Walpole & Myers 8.6, Dixon & Massey 7-2
Two-Sample t-Test	3.3.1	Walpole & Myers 8.4, Dixon & Massey 8-4
Two-Sample Test for Proportions	3.3.2	Walpole & Myers 8.7, Dixon & Massey 13-6
Wilcoxon Rank Sum Test	3.3.3	Walpole & Myers 14.4, Dixon & Massey 17-4
Shapiro Wilk W Test	4.2.2	Gilbert 12.3.1, EPA (1992a) 1.1.4
Filliben's Statistic	4.2.3	EPA (1992a) 1.1.6
Coefficient of Variation Test	4.2.4	EPA (1992a) 4.2.2
Skewness and Kurtosis Tests	4.2.5	Madansky 1.4
Geary’s Test	4.2.6	Madansky 1.3
Studentized Range Test	4.2.6	Madansky 1.3
Goodness-of-Fit Tests	4.2.7	Walpole & Myers 8.9, Dixon & Massey 13-4
Test of a Correlation Coefficient	4.3.2	Walpole & Myers 9.9
Extreme Value Test	4.4.3	Dixon & Massey 16-3
Discordance Test	4.4.4	EPA (1992a) 9-2
Rosner's Test	4.4.5	Gilbert 15.3
Walsh's Test	4.4.6	Madansky 4.2
Confidence Intervals for a Variance	4.5.1	Walpole & Myers 7.8, Dixon & Massey 7-3
F-Test	4.5.2	Walpole & Myers 8.8, Dixon & Massey 8-3
Bartlett’s Test	4.5.3	Walpole & Myers 11.3, Dixon & Massey 15-5

Table B-1: Selected References from Primary and Introductory Textbooks (Continued)

Subject	Section	Source (with Chapter)
Levene's Test	4.5.4	EPA (1992a) 1.2
Trimmed Mean & Winsorization	4.7.2	Dixon & Massey 16-4
Cohen's Adjustment	4.7.2	EPA (1992a) 8.1.3

Table B-2: Secondary References

Subject	Section	Source
Profiles	2.3.7	Wang (1978)
Stars	2.3.7	Siegel, Goldwyn and Friedman (1971)
Glyphs	2.3.7	Kleiner and Hartigen (1981)
Parallel Coordinate Plots	2.3.7	Wegman (1990)
W-test	4.2.2	Shapiro and Wilk (1965)
D'Agostino's Test	4.2.3	D'Agostino (1971)
Royston's Test	4.2.3	Royston (1982)
Filliben's Statistic	4.2.3	Filliben (1975)
Geary's Test	4.2.6	Geary (1935, 1947)
Studentized range test	4.2.6	David, Hartley, and Pearson (1954)
Kolmogorov-Smirnoff Test	4.2.7	Conover (1980)
Lilliefors K-S Test	4.2.7	Lilliefors (1967, 1969)
Sen's Slope Estimator	4.3.3	Sen (1968a, 1968b)
Extreme Value Test	4.4.3	Dixon (1953)
Discordance Test	4.4.4	Grubbs (1969)
Rosner's Test	4.4.5	Rosner (1975)
Walsh's Tests	4.4.6	Walsh (1958 and 1950)
Bartlett's Test	4.5.3	Dixon and Massey (1983)
Cohen's Method	4.7.2	Cohen (1959)