EM 200-1-16
31 May 2013

US Army Corps
of Engineers®

ENVIRONMENTAL QUALITY

ENVIRONMENTAL STATISTICS

ENGINEER MANUAL

AVAILABILITY

Electronic copies of this and other U.S. Army Corps of Engineers (USACE) publications are available on the Internet at http://140.194.76.129/publications/. This site is the only repository for all official USACE engineer regulations, circulars, manuals, and other documents originating from HQUSACE. Publications are provided in portable document format (PDF).

This document is intended solely as guidance. The statutory provisions and promulgated regulations described in this document contain legally binding requirements. This document is not a legally enforceable regulation itself, nor does it alter or substitute for those legal provisions and regulations it describes. Thus, it does not impose any legally binding requirements. This guidance does not confer legal rights or impose legal obligations upon any member of the public.

While every effort has been made to ensure the accuracy of the discussion in this document, the obligations of the regulated community are determined by statutes, regulations, or other legally binding requirements. In the event of a conflict between the discussion in this document and any applicable statute or regulation, this document would not be controlling.

This document may not apply to a particular situation based upon site-specific circumstances. USACE retains the discretion to adopt approaches on a case-by-case basis that differ from those described in this guidance where appropriate and legally consistent.

This document may be revised periodically without public notice.

DEPARTMENT OF THE ARMY
U.S. Army Corps of Engineers
Washington, D.C. 20314-1000

EM 200-1-16

CEMP-CE

Manual
No. 200-1-16

31 May 2013

Environmental Quality
ENVIRONMENTAL STATISTICS

1. Purpose. The primary purpose of this Engineer Manual (EM) is to provide practical guidance for statistical evaluations of environmental chemical data to ultimately improve the quality of decisions. The foundation of Corps of Engineers environmental work is the Environmental Operating Principles as specified in ER 200-1-5. These seven tenets serve as guides and must be applied in all Corps business lines as we strive to achieve a sustainable environment

2. Applicability. This EM applies to all USACE commands having Civil Works and/or Military Programs hazardous, toxic, or radioactive waste (HTRW) project responsibilities.

3. References. References are provided in Appendix A.

4. Distribution Statement. Approved for public release, distribution is unlimited.

5. Discussion. This manual provides an overview of statistical methods that are applicable to the various life cycles of a typical environmental project. The manual explains basic statistical concepts and their application to environmental projects. The manual may be used as a desk-top reference, as it provides step-by-step instructions for conducting a variety of useful and common statistical tests for environmental data. However, it should be noted that the manual is not intended to replace statistical texts or electronic statistical software. It does not present derivations of statistical formulas or a comprehensive treatment of statistical concepts, but focuses on the application of select statistical methods.

FOR THE COMMANDER:

19 Appendices
(See Table of Contents)

C. DAVID TURNER
Colonel, Corps of Engineers
Chief of Staff

This manual supersedes EM 1110-1-4014 dated 31 January 2008.

DEPARTMENT OF THE ARMY
CEMP-CE U.S. Army Corps of Engineers
Washington, DC 20314-1000

Manual
No. 200-1-16 31 May 2013

Environmental Quality
ENVIRONMENTAL STATISTICS

Table of Contents

	Paragraph	Page
Chapter 1. Introduction		
Purpose	1.1	1-1
Applicability	1.2	1-1
Distribution Statement	1.3	1-1
References	1.4	1-1
Introduction	1.5	1-1
Technical Project Planning and the Project Life Cycle	1.6	1-3
Data Quality Objectives, Data Quality Indicators, and Measurement Quality Objectives	1.7	1-4
Statistics in Environmental Project Planning	1.8	1-10
Chapter 2. Preliminary Assessment/Site Investigation		
Section I. Preliminary Assessment		
Introduction	2.1	2-1
Case Study 1—Examining Historical Data Sets	2.2	2-1
Section II. Site Inspection		
Introduction	2.3	2-3
Sampling Design	2.4	2-4
Case Study 2—Judgmental Sampling, Oil/Water Separator	2.5	2-5
Case Study 3—Arsenic Contamination in Soil	2.6	2-6
General Review of Sample Size Determination	2.7	2-7
Case Study 4—Effect of Decision Confidence on Sample Number	2.8	2-8
Summary of Case Studies	2.9	2-9
Comparing On-site Data to Fixed Screening Criteria	2.10	2-9
Quantitative Statistical Evaluations	2.11	2-11
Chapter 3. Remedial Investigation/Feasibility Study		
Introduction	3.1	3-1
Section I. Site Characterization		
Introduction	3.2	3-2
Case Study 1—Hot-Spot Identification	3.3	3-3

Paragraph Page

Case Study 2—Using Geostatistics in Project Planning to Reduce Uncertainty and Cost 3.4 3-4

Section II. Background Comparisons
Introduction 3.5 3-5
Does Background Soil Differ From Site Soil? 3.6 3-7
Simple Background Comparison 3.7 3-9
Case Study 3—Comparison to Regional Background 3.8 3-8
Parametric and Nonparametric Tests 3.9 3-10
Case Study 4—Establishing and Comparing Background Concentrations to On-site Data 3.10 3-11
Upper Tolerance Limits 3.11 3-12
Case Study 5—Calculating Background UTLs for Metals 3.12 3-13
Extended Background Example 3.13 3-14

Section III. Risk Assessment
Introduction 3.14 3-22
Case Study 6—Refining Risk Assessment Assumptions 3.15 3-25
Case Study 7—Direction and Magnitude of Bias 3.16 3-26

Section IV. Probabilistic Risk Assessments Monte Carlo Simulations
Introduction 3.17 3-27
Case Study 8—Monte Carlo Simulation in Remedial Alternative Selection 3.18 3-28

Chapter 4. Remedial Design/Remedial Action

Introduction 4.1 4-1
Comparisons to ACLs and MCLs 4.2 4-1

Section I. Groundwater Monitoring and Optimization Trend Analysis
Introduction 4.3 4-2
Detection/Compliance Monitoring 4.4 4-3
Case Study 1—Groundwater Monitoring 4.5 4-4
Case Study 2—Shewart/CUSUM Monitoring 4.6 4-10
Optimization 4.7 4-12
Case Study 3—Trend Analysis and Correlation in Natural Attenuation Data 4.8 4-14

Section II. Applying Cleanup Levels
Introduction 4.9 4-15
Determining Not-to-Exceed Concentrations Using Non-Spatial Statistics 4.10 4-16

APPENDICES

Page

Appendix A—References A-1
Appendix B—Statistical Tables B-1
Appendix C—Sampling Strategies C-1
Appendix D—Descriptive Statistics D-1
Appendix E—Statistical Distributions E-1
Appendix F—Testing for Normality F-1
Appendix G—Detection Limits and Quantitation Limits G-1
Appendix H—Censored Data H-1
Appendix I—Identification and Handling of Outliers I-1
Appendix J—Graphical Tools J-1
Appendix K—Intervals and Limits K-1
Appendix L—Hypothesis Testing Simple Cases L-1
Appendix M—Hypothesis Testing Two-Population and General Cases M-1
Appendix N—Hypothesis Testing Tests of Dispersion N-1
Appendix O—Measures of Correlation O-1
Appendix P—Comparing Laboratory and Field Data P-1
Appendix Q—Trend Analysis Q-1
Appendix R—Geostatistics R-1
Appendix S—Geochemical Trend Analysis S-1
Glossary Glossary-1

LIST OF TABLES

Table 1-1 Project Phase Crosswalk between CERCLA and RCRA 1-2
Table 1-2 Crosswalk Between the TPP and DQO Processes 1-5
Table 1-3 Variability in Environmental Studies 1-8
Table 2-1 Site Screening Data Table 2-13
Table 3-1 Background Population Comparison 3-11
Table 3-2 Calculation of Background UTL_S 3-13
Table 3-3 Analytical Results for Chromium at Site A and Background Locations 3-15
Table 3-4 Results of the Shapiro-Wilk Test of Normality and Lognormality for Chromium Surface Soil at Site A and Background 3-21
Table 3-5 Results of the Studentized Range Test of Normality and Lognormality for Chromium Surface Soil at Site A and Background 3-21
Table 4-1 Groundwater Monitoring Data for Case Study 1 4-5
Table 4-2 Eight Most Recent TCE Measurements in B-37 4-11
Table 4-3 TCE Measurements and Shewart/CUSUM Calculations 4-11
Table 4-4 Attenuation Data 4-14
Table B-1 Binomial Distribution B-1
Table B-2 Percentiles of the Chi-Square Distribution B-10
Table B-3 Values of the Parameter λ for Cohen's Estimates B-11

Page

Table B-4 Critical Values of *D* for the Discordance Test B-12
Table B-5 Critical Values of Dixon's Test (Extreme Value Test).......................................B-13
Table B-6 Critical Values for Duncan's Multiple Range Test...B-14
Table B-7 Percentiles of the F Distribution B-15
Table B-8 H-statistic for Confidence Limit on a Lognormal Mean B-16
Table B-9 Quantiles of D'Agostino's Test for Normality B-19
Table B-10 Probabilities for the Small-Sample Mann-Kendall Test for Trend......................B-19
Table B-11 Confidence Levels for Nonparametric Prediction Limits....................................B-21
Table B-12 Nonparametric Confidence Intervals on a Proportion B-24
Table B-13 Factors for Calculating Normal Distribution On-Sided Tolerance Limits B-25
Table B-14 Factors for Calculating Normal Distribution Two-Sided
Tolerance Intervals...B-27
Table B-15 Standard Normal Distribution...B-28
Table B-16 Poisson Probabilities...B-30
Table B-17 Critical Values for the Rank-Sum Test...B-34
Table B-18 Approximate Critical Values (λ_ι) for Rosner's Test..B-35
Table B-19 Coefficients for the Shapiro-Wilk W Test for Normality....................................B-37
Table B-20 Quantities of W_α of the Shapiro-Wilk Test for Normality....................................B-39
Table B-21 Critical Values for the Studentized Range Test...B-40
Table B-22 Percentage Points of the Studentized Range..B-41
Table B-23 Critical Values of Student's t-Distribution B-43
Table B-24 Quantiles of the Wilcoxon Signed Rank Test..B-44
Table B-25 Modified Quantile Test Critical Numbers Level of Significance (α)..................B-45
Table B-26 Dunnett's Test (One-Tailed) Total Number of Investigate Groups (K-1)............B-49
Table B-27 Upper Tail Critical Values for the *F*-Max Test B-50
Table B-28 Power of ANOVA for K (varied # of groups and significance levels)................B-51
Table C-1 Hot-Spot Grid Spacing..C-6
Table C-2 Comparison of Random Sampling Method Results...C-11
Table C-3 Objectives of Composite Sampling-Fundamental Cases.....................................C-12
Table C-4 Criteria for Judging Benefits of Composite Sampling...C-13
Table C-5 Compositing Methods...C-14
Table F-1 Methods Available to Verify the Assumption of Normality F-2
Table H-1 Guidelines for Analyzing Data with NDs... H-6
Table H-2 Case Study, Sodium in Surface Soil: Summary Statistics Using Various
Substitution Methods for Proxy Values.. H-11
Table H-3 Benzene Concentrations, the DL, and Natural Log of the Concentrations.......... H-14
Table H-4 Groundwater Concentrations for Benzene at Site A for Well MW03................. H-21
Table I-1 Background Surface Soil Data for Antimony..I-11
Table I-2 Summary Statistics for Antimony Background Surface Soil Data........................I-12
Table I-3 Recommendations for Selecting a Statistical Test for Outliers I-13
Table J-1 Number of Observations within each Interval...J-4
Table J-2 Quantile Plot Data..J-11

Page

Table J-3 Normal Probability Data J-14
Table J-4 Spatial Data J-26
Table J-5 Arsenic and Chromium Concentrations in Subsurface Soil at Site A J-31
Table K-1 UCL Method Flow Chart K-4
Table K-2 Example Data K-8
Table K-3 Example Compliance Well Data K-16
Table K-4 Recommended Methods of Computation of a 95% UCL for the Unknown Mean of a Lognormal Population K-18
Table K-5 Estimate of $\hat{\theta}_{(i)}$ for i = 1, 2 … *n* = 36 .... K-32
Table K-6 Background Wells K-43
Table L-1 Commonly Used Statements of Statistical Hypotheses L-5
Table L-2 Conclusions Associated with Any Statistical Hypothesis Test L-6
Table L-3 Relationship Between Hypothesis Tests and Confidence Intervals L-7
Table L-4 Example of L-6.1.3 Data L-17
Table L-5 Data for Example L-6.2.2 .... L-22
Table L-6 Analytical Results from Samples Collected at the Site in Example L-6.3.2 .... L-26
Table L-7 Analytical Results from Samples Collected at the Site in Example L-6.4.3 .... L-28
Table L-8 All Analytical Results from Samples Collected from Both Sampling Events L-31
Table L-9 Analytical Results from Samples Collected for the Pilot Study and Independent Pair Wise Averages L-33
Table L-10 Analytical Results from Samples Collected at the Site for Example L-7.3 .... L-38
Table M-1 TCE Concentrations (*mg*/L at Monitoring Wells Before and After Treatment M-10
Table M-2 Example M-4.14 Pb Concentrations M-17
Table M-3 Example M-4.1.5.2 Pb Concentrations M-19
Table M-4 TCE Concentrations (mg/L) at Monitoring Wells Before and After Treatment for Examples M-4.1.6.2 .... M-24
Table M-5 Nickel Surface Soil Concentrations for Example M-4.2.3 .... M-26
Table M-6 Manganese (Mn) Groundwater Concentrations to be Compared Among the Wells at Site A M-31
Table M-7 Summary Statistics for Mn by Well M-32
Table M-8 Lead Concentrations for Example M-5.2.3 .... M-35
Table M-9 Summary of Multiple Comparison Tests M-37
Table M-10 Results of Example M-6.2.2 .... M-39
Table M-11 List of the Means in Order for Example M-6.2.2 .... M-40
Table M-12 Means with the Same Letter are not Significantly Different at $\alpha = 0.05$ in Example M-6.3.2 .... M-42
Table M-13 Results from Example M-6.4.2 .... M-43
Table M-14 Means with the Same Letter are not Significantly Different at $\alpha = 0.05$ M-44
Table M-15 Means with the same Letter are not Significantly Different at Significance of $\alpha = 0.05$ M-47
Table M-16 Data for Example M-6.6.2 .... M-50
Table N-1 Data for Example N-5.2 .... N-10

Page

Table O-1 Matrix of Differences for Kendall's τ O-14
Table O-2 Matrix of Differences for Kendall's τ for Data in O-2.7.3.1 .... O-16
Table Q-1 Monthly Average Temperature (°F) in Austin, Texas, from 1995 through 1998 .... Q-3
Table Q-2 Adjusted Monthly Average Temperature (°F) in Austin, Texas from 1995 Through 1998 .... Q-5
Table Q-3 Basic Mann-Kendall Trend Test with a Single Measurement at Each Time Plot Q-13
Table Q-4 "Upper Triangular" Data for Basic Mann-Kendall Trend Test with a Single Measurement at Each Time Point-Data Table Q-15
Table Q-5 Data for Example Q-3.5.5 .... Q-16
Table Q-6 Benzene Data for Example Q-3.3.9 .... Q-22
Table Q-7 Pairwise Slopes Data Table Q-25
Table Q-8a Historical Data for Upper Control Limit in Example Q-4.2.3 .... Q-31
Table Q-8b Current Data to Use to Compare to Control Limit in Example Q-4.2.3 .... Q-31
Table Q-9 Results for Combined Shewart/CUSUM Control Chart Q-35

LIST OF FIGURES

Figure 1-1 Project Planning Life Cycle 1-12
Figure 1-2 Guidance Document Life Cycle 1-13
Figure 1-3 Balance between Resources and Certainty 1-13
Figure 1-4 Project Planning Triad 1-14
Figure 3-1 Initial Sampling Grid and Proposed New Samples 3-5
Figure 3-2 Samples Required After Geostatistical Analysis 3-6
Figure 3-3 Chromium in Site A 3-17
Figure 3-4 Chromium in Background 3-18
Figure 4-1 1989 EPA Decision Tree for Groundwater Monitoring 4-6
Figure 4-2 Statistical Decision Tree with Options for Groundwater Monitoring-Part 1 ... 4-7
Figure 4-3 Statistical Decision Tree with Options for Groundwater Monitoring-Part 2 ... 4-8
Figure 4-4 Statistical Decision Tree with Options for Groundwater Monitoring-Part 3 ... 4-9
Figure 4-5 Statistical Decision Tree with Options for Groundwater Monitoring-Part 4. 4-10
Figure 4-6 Shewart/CUSUM Control Chart, Well B-37 4-12
Figure 4-7 PCE Concentration Versus Distance 4-15
Figure 4-8 Geochemical Parameters Versus Distance from Source 4-15
Figure C-1 Elliptical Hotspot C-7
Figure C-2 Population Grid with Initial and Follow-up Samples and Areas of interest C-18
Figure C-3 Adaptive Sampling Plan Flow Chart C-21
Figure E-1 Probability Density Function E-2
Figure E-2 Normal Distribution E-3
Figure E-3 Standard Normal (Z) Curve E-3
Figure E-4 95th Percentile of the Standard Normal Distribution E-4

Page

Figure E-5 Comparison of Students t-Distribution with Standard Normal Distribution E-8
Figure E-6 Lognormal Distributions E-9
Figure G-1 Noise in GC Baseline G-2
Figure G-2 "Type I DL" (L_c) and "Type II DL" (L_p) G-3
Figure H-1 Example of Selecting between Atchison's Method and Cohen's Method......... H-19
Figure I-1 Box-and-Whisker Plot for Antimony........I-11
Figure J-1 Frequency Plot: Normal Data........J-2
Figure J-2 Histogram: Lognormal Data........J-3
Figure J-3 Frequency Plot........J-5
Figure J-4 Histogram J-5
Figure J-5 Examples of Box-and-Whiskers Plots........J-7
Figure J-6 Box-and Whiskers Plot........J-9
Figure J-7 Examples of Quantile Plots J-10
Figure J-8 Example of a Quantile Plot J-12
Figure J-9 Examples of Normal Probability Plots........J-13
Figure J-10 Example of a Normal Probability Plot J-14
Figure J-11 Empirical *q*-*q* Plot J-17
Figure J-12 Examples of Time Plots J-19
Figure J-13 Example Time Plot........J-21
Figure J-14 Correlogram for Data in Paragraph J-7-4.3........J-21
Figure J-15 Posting Plot........J-25
Figure J-16 Symbol Plot J-25
Figure J-17 Concentrations of Napthalene at Site A Wells........J-28
Figure J-18 Graphical Representations of Multiple Variables J-28
Figure J-19 Example of a Scatter Plot........J-30
Figure J-20 Scatter Plot Matrix........J-32
Figure J-21 Napthalene Concentrations at Site A Wells: Side-by-side Scatter Plots.............J-33
Figure J-22 Contour Plot Drawn by Linear Interpolation J-36
Figure J-23 Contour Plot Drawn by Linear Interpretation Incorporating Site Knowledge.....J-37
Figure J-24 Contour Plot Drawn by Modeling Software Using IDW........J-24
Figure L-1 Decision Errors Associated with a Hypothesis Test L-9
Figure L-2 Power Curve for the One-sample τ-Test for Simple Random Sampling L-18
Figure L-3 Power Curve for the One-sample τ-Test for Stratified Sampling L-25
Figure L-4 Histogram Plot of Wilcoxon Signed Rank Test for Random Samples L-29
Figure L-5 Histogram Plot of Wilcoxon Signed Rant Test for Large Random Samples......L-32
Figure M-1 Estimated Power Performance Curve........M-7
Figure O-1 Scatter Plot for Chromium and Lead........ O-7
Figure O-2 Scatter Plot with Outlier........ O-8
Figure Q-1 Monthly Average Temperature (°F) in Austin, Texas from 1995 through 1998........ Q-4
Figure Q-2 Adjusted Monthly Average Temperature (°F) in Austin, Texas from 1995 through 1998........ Q-6
Figure Q-3 Residuals Versus the Independent Variable........ Q-9

Page

Figure Q-4 Trend for Benzene in Groundwater (small sample size) Q-14
Figure Q-5 Trend for Benzene in Groundwater (large sample size) Q-17
Figure Q-6 Time Plot for Seasonal Series with Decreasing Trend Q-23
Figure Q-7 The Correlogram Q-31
Figure Q-8 Historical Data (1998-2000) and 2001 Data with Shewart Control Limit for Benzene (SW8260B) in Groundwater at Site A, MW-01 .... Q-32
Figure Q-9 Combined Shewart-CUSUM Control Chart Q-36
Figure R-1 Types of Semivariograms R-4
Figure R-2 Comparison of Krige-Interpolated Benso(α)pyrene Concentrations before and after Treatment by In Situ Ozonation R-7
Figure R-3 Surface Topography with Benzo(α)pyrene Concentrations and Kriged Isovolume R-9
Figure R-4 Two Sets of Experimental Variograms Used to Define the Krige Variogram Model R-11
Figure S-1. Scatter Plot of Pb and Fe S-3
Figure S-2 Log Scale Cu-Fe Scatter Plots of Site and Background Soil Samples S-4
Figure S-3 Log Scale PB-Mn Scatter Plot for Site and Background Soil Samples S-5
Figure S-4 Log Scale Cu-Pb Scatter Plot of Background and Site Soil Soils S-5
Figure S-5 As-Fe Scatter Plot with a Large Amount of Scatter S-6
Figure S-6 Misidentified Trends for the Scatter Plot in Figure S-5 .... S-6
Figure S-7 Scatter Plot of As and Logarithm of As/Fe Using the Data Set Plotted in Figure S-5 .... S-7
Figure S-8 Al-Fe Log-scale Scatter Plot for a Set of Groundwater Monitoring Wells S-8
Figure S-9 Log-scale As-Fe Scatter Plot Using Fe Groundwater Date for Figure S-8 .... S-8
Figure S-10 Log-scale Scatter Plot of Filtered and Unfiltered Groundwater Analyzed for Cr S-9
Figure S-11 Probability Plot of $Y = (C_M/C_X)$ Site When a Portion of the Study Area has been Heavily Impacted by Anthropogenic Contamination S-10
Figure S-12 Probability Plot of $Y = (C_M/C_X)$ *Site* for Three Different Populations S-11
Figure S-13 Probability Plot of $Y = (C_M/C_X)$ *Site* When a Portion of the Study Area has been Slightly Impacted by Anthropogenic Contamination S-12
Figure S-14. Scatter Plot for Censored Data S-14

CHAPTER 1

Introduction

1.1. Purpose. This document is intended to serve as a guide to project team members for the use of statistics in environmental decision-making. The foundation of Corps of Engineers environmental work is the Environmental Operating Principles as specified in ER 200-1-5. These seven tenets serve as guides and must be applied in all Corps business lines as we strive to achieve a sustainable environment.

1.2. Applicability. The U.S. Army Corps of Engineers (USACE) developed this document within the broader scope of Technical Project Planning (TPP), recognizing that understanding statistical evaluations can improve project planning and implementation at hazardous, toxic, and radioactive waste (HTRW) sites.

1.3. Distribution Statement. Approved for public release; distribution unlimited.

1.4. References. References are contained in Appendix A.

1.5. Introduction. This Manual's primary objective is to improve a decision-maker's understanding of common environmental statistical evaluations. The applicability of statistical tests and considerations is presented in the context of a typical environmental project life cycle. This document should serve as a first step in explaining statistical concepts and their application at HTRW sites. It is not intended to replace more robust statistical texts or electronic statistical software.

1.5.1. Statistics are applicable to environmental projects throughout their entire life cycle and yield defensible, cost-effective solutions to environmental questions. Statistics can be used to guide the selection of sampling locations, analyze large data sets, and verify that project objectives have been met. Statistics are of particular importance for quantifying the power and limitations of environmental data, specifically because these data are usually limited. It is not possible to collect and analyze every bit of an environmental medium (for example, soil, sediment, groundwater, or surface water) at a site; instead, a set of sample data is used to characterize the environmental medium as a whole.

1.5.2. This Manual is organized into four major Chapters, each associated with a stage in a typical Superfund project life cycle. These Chapters are supported by Appendices that provide detailed statistical or technical explanations of concepts or techniques used within the main sections.

1.5.3. Statistical terms unfamiliar to some readers may be used in the four main chapters. When used for the first time, these terms will be printed in italics and footnoted. The footnote will direct the reader to the appropriate Appendix for a detailed explanation of the term. To demonstrate the types of statistical concepts necessary for the planning stages of environmental projects, concepts are presented in the context of Comprehensive Emergency Response, Compensation, and Liability Act (CERCLA) projects. The material is applicable to Resource Conservation and Recovery Act (RCRA) projects as well. The steps involved in the two programs are similar except for the use of different terminology and the applicable regulations. Table 1-1 presents a terminology crosswalk for the stages of CERCLA and RCRA investigations.

1.5.4. In the following Chapters of this document, major stages that require data gathering and evaluation are presented, and to the extent that statistical processes are applicable, examples are provided from case studies illustrating the application of those statistical processes. Some statistical elements may apply in more than one phase of the project life cycle. The Appendices provide detailed instructions on implementing the statistical processes.

1.5.5. The CERCLA project life cycle is not always linear. As information regarding a given site is gathered, additional questions may be raised about a previously unrecognized threat to human health or the environment. In that case, the process can repeat in whole or in part, creating a series of loops to previous portions of the cycle. In addition, at any point in the process, emergency activities (e.g., “time critical” remedial actions) may occur at earlier or later times in the cycle. Finally, the process can terminate at the end of any given phase in a “no further action” determination.

Table 1-1. Project Phase Crosswalk Between CERCLA and RCRA

CERCLA Project Phase	RCRA Project Phase
Discovery and Notification	Permit Application
Preliminary Assessment	RCRA Facility Assessment
Site Investigation	Site Inspection
Hazard Ranking	Administrative Order
Remedial Investigation	RCRA Facility Investigation
Feasibility Study	Corrective Measures Study
Proposed Plan	Statement of Basis
Record of Decision	RCRA Permit
Remedial Design	Remedy Design
Remedial Action	Corrective Measures Implementation
Five Year Review	Monitoring/Annual Report
Closeout	Closure

1.5.6. The remedial action process under CERCLA is necessarily iterative and the same statistical tools can be employed repeatedly to address the original problem or newly identified issues at the site. For purposes of this text, however, we will assume a linear progression through an idealized project life cycle, consistent with the instructions contained in EM 200-1-2.

1.5.7. In the Technical Project Planning Process, the user is encouraged to identify the appropriate project phase for a given segment of work, then reference matching portions of this Manual for statistical guidance and methods appropriate to that phase.

1.6. Technical Project Planning and the Project Life Cycle. EPA QA/G-4 states, “EPA Order 5360.1 A2 [requires that] all EPA organizations (and organizations with extramural agreements with EPA) follow a systematic planning process to develop acceptance or performance criteria for the collection, evaluation, or use of environmental data.” Similarly, ER 5-1-11 states, “Requirements for quality must be addressed during the planning phase of a project’s life cycle, rather than waiting until the review or inspection stage.” Thus, a systematic planning process of some sort is required for all HTRW projects involving the collection of data.

1.6.1. The EPA approach to systematic planning is described in detail in EPA QA/G-4 and is called the Data Quality Objectives (DQO) process. It is a seven-step process, which has as its goal the design of legally and scientifically defensible sampling strategies. The DQO guidance generally assumes that decision-making requires a probabilistic approach. Fundamental to the DQO process is identifying some statistic describing an environmental site that is compared via a statistical process to either a fixed threshold or risk-based value, or a statistical comparison of some descriptive measure of data for two or more variables. The DQO process also incorporates statistical tools for estimating such things as the number of samples required to measure a site characteristic, spacing of sampling locations, and frequency of sampling. This permits data users to make decisions with specific degrees of statistical confidence.

1.6.2. The USACE TPP process is broader in scope, with the EPA’s DQO process as one step within it, to the extent that probabilistic decision-making is appropriate to the goals of the project. The intent of the TPP process is to “get to closure” and to provide documentation of project decisions and project performance. The TPP process is useful for all sites, regardless of whether probabilistic decision-making is involved. It is highly flexible and promotes an approach that balances the size and complexity of a given site or problem with the level of effort involved in the planning process.

1.6.3. As described in EM 200-1-2, there are four phases to the TPP process, as follows.

1.6.3.1. Identify the Current Project Phase. The project manager establishes a project team to encompass all of the perspectives and skills required to take the project from beginning to end. The project manager briefs the team on client goals and existing site information and develops a conceptual model for the site. A broad, overall approach to the work is agreed upon, including an assessment of the most likely remedies or outcomes for the site. The work is broken down into clearly defined executable stages and the current stage of work is identified.

1.6.3.2. Determine Data Needs. Allowing all perspectives to be addressed, the team identifies the data required for each data user type (e.g., hydrogeologic, chemical, health and safety, risk assessment, engineering, etc.). The team reviews sources of existing information for availability, quality, and applicability to the current stage of work, and identifies data gaps that only new data can fill.

1.6.3.3. Develop Data Collection Options. With their respective needs defined, the team members decide on the best approach to obtain the required data. Usually, the team assesses a number of differing approaches and selects the approach that provides all of the requisite data with the best balance of available resources, measurement quality, and client risk tolerance. The TPP process clearly defines three data collection options: basic, optimum, and excessive. A basic sampling approach provides data applicable only to the current stage of work, whereas an optimum approach addresses both current data needs and anticipated future needs as well. An approach not focused on the specific data required to "get to closure" is excessive and should be avoided.

1.6.3.4. Finalize the Data Collection Program. At this point, the team encourages clients, regulators, the public, and in some cases other parties, to take part in the decision-making process. Specific DQO statements are prepared for each data user and data type and, to the extent that probabilistic decision-making is appropriate, the EPA's DQO guidance document (EPA QA/G-4) is used and applied to these statements. From these DQO statements, scopes of work and other project controlling documents (PCDs) such as work plans, quality assurance (QA)/quality control (QC) plans, field sampling plans (FSPs), etc., are derived and cost estimates generated.

1.6.4. Table 1-2 provides a crosswalk between the EPA DQO Process and the USACE TPP process.

1.6.5. Failure to apply, or to apply properly, the TPP process can result in a variety of negative consequences. Failure to properly plan for data collection may require more time and money to implement the work. Lack of planning may extend the time it takes to validate work because both objectives and verification methods may be unclear. Poor planning may create the need for extensive rework or remobilization. Finally, lack of advance planning can cause increases in legal risk to the client and to the USACE by increasing the potential for decision error. On the other hand, too great an emphasis on planning extends the planning cycle and the checking cycle, depleting the available resources.

1.7. Data Quality Objectives, Data Quality Indicators, and Measurement Quality Objectives. This paragraph provides a conceptual understanding of DQOs in the context of project planning for environmental investigations and remediations. The terminology is less important than the underlying concepts that support the decision-making process, as long as all parties possess a common understanding of that process. Project planners derive DQOs from scientific objectives, as well as social and economic objectives and the regulatory objectives of the environmental program under which the project is implemented. DQOs are technical, goal-oriented, qualitative, and quantitative statements derived from the planning process that clarify study objectives, define the appropriate type of data, and specify tolerable levels of potential decision error. The DQO process typically uses statistics and is the basis for establishing the quality and quantity of data needed to support decisions. The DQO process does not establish specifications for data quality—called measurement quality objectives (MQOs)—or the mechanisms for measuring conformance to those specifications—called data quality indicators (DQIs). MQOs and DQIs are discussed in additional detail below.

Table 1-2. Crosswalk Between the TPP and DQO Processes

EPA's DQO Process		**USACE TPP Process**			
		Phase I	**Phase II**	**Phase III**	**Phase IV**
Step 1 **State the Problem**		Identify the Current Project	Determine Data Needs	Develop Data Collection Options	Finalize Data Collection Program
Step 2 **Identify the Decision**					
Step 3 **Identify Inputs to the Decision**				Develop Data Collection Options	
Step 4 **Define the Study Boundaries**					
Step 5 **Develop a Decision Rule**		Identify the Current Project			
Step 6 **Specify Limits on Decision Error**					
Step 7 **Optimize the Design**					Finalize Data Collection Program

1.7.1. Data Quality. Data quality depends on the integrity of each element in a series of events. It is critical to collect samples that are representative of the features of the environmental population being investigated in the study area. Representativeness depends on factors such as sample frequency, location, time of collection, and the nature of the sampled medium. Pre-testing factors include sample containerization, preservation, transportation, and storage. Sample analysis factors generally include sample homogenization, sub-sampling, sample preparation (such as extraction and cleanup), as well as the instrumental analysis of the sample. The final steps of the process include data generation, reduction, and review.

1.7.1.1. Historically, attention has been focused primarily on the analytical component of data quality rather than on "total measurement system quality." Environmental decision-makers and practitioners tend to assume that data quality is primarily determined by the analytical methodology. For example, as fixed laboratory methods tend to be superior to field methods in terms of analytical uncertainty, data produced from field methods have been viewed to be too uncertain to support critical project decisions. However, defensible decisions are possible only when data quality encompasses total uncertainty rather than the uncertainty associated with only the analytical portion of the investigation. The value of data is limited less by the analytical procedures than by the quality of the sampling design[*] and the inherent variability of the environmental population of interest or condition being measured (the "field" component of variability). Because analytical uncertainty is typically small relative to field uncertainty, data quality usually depends more on sampling design than the quality of the individual test methods.

1.7.1.2. Table 1-3 summarizes sources or components of variability for environmental studies and how they are measured and controlled.

1.7.1.3. Regulators have also historically insisted on adhering to pre-approved analytical methods because of a perception that this ensures defensible data and that definitive data will be produced when EPA-approved analytical methods and QA/QC requirements are used. Though adequate data quality is often achieved using EPA-approved analytical methods, they are insufficient to ensure data of high quality. Efforts to improve data quality have primarily focused upon increasing laboratory oversight, rather than on developing mechanisms to manage the largest sources of uncertainty in data, which are issues related to sampling. Furthermore, prescriptive methods are scientifically feasible only when the sample matrices do not vary in any manner that will affect the reliability of the analyses. As all analytical methods are potentially subject to chemical and physical interferences, given the variability and complexity of environmental matrices, it is unlikely that "one-size-fits-all" analytical methodologies are viable for all projects.

1.7.1.4. The EPA has recently clarified its intended meaning of the term "data quality" in its broadest sense by defining it as "the totality of features and characteristics of data that bear on its ability to meet the stated or implied needs and expectations of the client." One must know how a data set is to be used to establish a relevant benchmark for judging whether the data quality is adequate. Linking data quality directly to their intended use provides a

[*] Appendix C.

firm foundation for building a vocabulary that distinguishes the individual components of overall data quality.

1.7.2. Data Quality Indicators. DQIs are qualitative and quantitative descriptions of data quality attributes: the various properties of analytical data historically expressed as precision, accuracy, representativeness, comparability, and completeness. Collectively, these factors are called the PARCC parameters. These are discussed in detail in EPA guidance documentation. Because it is evaluated at the same time, an additional parameter often combined with the PARCC parameters is sensitivity, which is the ability of an analytical method or technology to reliably identify a compound in the sample medium.

1.7.2.2. Completeness has been assigned an arbitrary goal of 80 to 100% based on the premise that decisions are still possible if a limited portion of the data are discarded (for example, because of quality control problems). However, the goal is based primarily on practical experience and is not mathematically based. Completeness should be evaluated in the context of project objectives.

1.7.2.3. In addition to these, selectivity is also a data quality indicator. “Selectivity” is the ability of an analytical method to identify the analyte of concern, e.g., the existence of other analytes in a sample or other interferences may mask the presence of the target analyte.

1.7.2.4. There may be more than one DQI for a single data quality attribute. For example, sensitivity is generally thought of in terms of detection, quantitation, or reporting limits, i.e., the lowest value that an analytical method can reliably detect or report. However, another important element of sensitivity is discrimination, the ability to distinguish between values to a given degree of precision. In other words, can the method tell the difference between values of 1 and 2 units, or only differences between 10 and 20 units? When developing DQIs, it is important to define them in terms of all the important attributes and assign specific numeric values to them as often as practicable.

1.7.3. Measurement Quality Objectives. MQOs are project-specific values assigned to DQIs derived from project-specific DQOs. MQOs are acceptance criteria for the DQIs and are derived by considering the level of measurement system performance needed to actually achieve project goals. MQOs are not intended to be technology - or method-specific. As with DQOs, MQOs specify what the level of data performance should be, but not how that level of data performance is to be achieved. A large part of the variability in environmental data stems from sampling considerations. MQOs should balance the relative contributions from analytical uncertainties and from sampling uncertainties. In many environmental media, matrix heterogeneity causes sampling variability to overwhelm analytical variability. Historically, the term MQO was restricted to the analytical side of the measurement process, but the broader concept of DQO (or decision confidence objectives) requires that sampling considerations be included. The importance of including both the sampling and analytical component of MQOs when assessing overall data quality cannot be overemphasized.

Table 1-3. Variability in Environmental Studies

Source of Variability	Measurement Method	Control Methods
Analytical Variability		
Analytical instrumentation	Replicate measurements of instrumental standards (most common for inorganic analysis)	Regular preventive maintenance
Analytical method	Duplicate analytical spikes, lab-blind field duplicate samples	Use of standard methods documented as standard operating procedures; control of standards and reagents; control of instrument conditions
Sample preparation method	Duplicate control samples and matrix spike/matrix spike duplicates	Use of standard methods documented as standard operating procedures; control of standards and reagents; regular, close supervision
Analyst	Analyst demonstration of capability, blank spikes/performance evaluation (PE) samples	Inter-laboratory comparison studies; internal PE and auditing programs; analyst training; regular, close supervision
Field Variability		
Sampling equipment	Field blanks	Routine inspection and preventive maintenance; decontamination; selection of appropriate equipment for representative samples
Sampling method	Method-specific standard deviation of field duplicate results	Selection of appropriate methods for representative samples
Sampler	Inter- and intra sampler standard deviation of field replicate results	Independent auditing program; training; regular, close supervision
Matrix heterogeneity	Field duplicates or replicates, matrix specific standard deviation of field replicates, matrix spike duplicates	Effective field mixing of sample components; compositing
Sample selection	Site-wide or stratum-specific standard deviation of field replicate results	Representative sampling plan; sufficient number of samples; statistically-based sampling design

Note: Duplicates are separate aliquots of the same sample; replicates are a second sample from the same location.

1.7.4. Relationships Among Decision Goals, DQOs, MQOs, and QC Protocols. During project planning, there should be a logical conceptual progression in the development of decision goals, DQOs, MQOs, and QC acceptance criteria. However, in practice, this will be a non-linear process.

1.7.4.1. As project planning develops, the following should be clearly presented:

1.7.4.1.1. General decision goals.

1.7.4.1.2. Technically expressed project goals (DQOs), and decision rules that will guide project decision-making.

1.7.4.1.3. Tolerable uncertainties for decisions.

1.7.4.1.4. Uncertainties that create decision errors.

1.7.4.1.5. Strategies for managing the uncertainties to achieve the desired tolerances for decision errors.

1.7.4.2. In the beginning of the project, program managers often set broad, non-technical goals. The next step is to translate these broad, non-technical goals into more technically oriented goals that can address specific considerations such as the following.

1.7.4.2.1. Regulations—what are the applicable environmental regulations?

1.7.4.2.2. Confidence in the outcome—how certain do we need to be by the end of the project that we have achieved goals such as risk reduction or regulatory compliance?

1.7.4.2.3. What are the constraints that need to be accommodated?

1.7.4.3. The next level of technical detail for data collection involves identifying DQIs and assigning to them project-specific MQOs that will be needed to achieve the project DQOs. At this point, the project team begins to consider in detail the options available for acquiring the needed measurements and selecting those that best meet the needs of the program. These decisions are documented in sampling and QC plans that specify the controls that will be used to ensure that MQOs are met and that any deviations are appropriately addressed.

1.7.4.4. Because sampling design and analytical strategy interact to influence the statistical confidence in final decisions, interaction among a statistician, a sampling expert, and an analytical chemist is critical for selecting a final strategy that can achieve project goals cost-effectively. The statistician is concerned with managing the overall variability of data, and with interpreting data with respect to the decisions being made. A statistician is a person having adequate familiarity with statistical concepts to correctly apply the required tests; this does not necessarily require a degree in statistics. The field sampling expert is responsible for implementing the sampling design while managing contributions to the sampling variability as actual sample locations are selected and as specimens are collected.

The chemist is responsible for managing components of variability that stem from the analytical effort.

1.7.4.5. In summary, the conceptual progression starts with the project-specific decision goals, and then moves from broader, higher-level goals to narrow, more technically detailed articulations of data quality needs. Project decisions are translated into project-specific DQOs; then into project-specific MQOs; then into technology/ method selection and development of a method-specific QC protocol that blends QA/QC needs of the technology with the QA/QC needs of the project. Then the process reverses. The data must be assessed against the project MQOs to document that data quality meets the decision-making needs of the project.

1.7.4.6. Figure 1-1 presents the life cycle in project planning. Figure 1-2 illustrates which guidance documents are useful in the planning phases of a project.

1.8. Statistics in Environmental Project Planning. The number of individual samples collected during a given study is called sample size and is generally designated by the statistic *n*. In order for decisions based on that sample to be meaningful in any scientific sense, the sample size has to be sufficiently large to account for the inherent variability in the characteristics measured. Sample size should be dependent on the variability in the measured condition but, in practice, is often limited by available resources.

1.8.1. A hypothetical illustration may be helpful in understanding this relationship. Let us suppose that a researcher wants to know the average concentration of a particular chemical constituent in the air of a sealed room. The constituent of interest is initially absent from the room and the researcher releases the chemical into the room from a port in the north wall of the room. Immediately after opening the port, a measurement taken along the south wall will not detect the presence of the chemical, while a sample taken adjacent to the port will display a high concentration. As the chemical disperses throughout the room via various physical processes, a single sample taken at any location in the room will not provide a representative value for the average concentration in the room as a whole. Even if a single sample were collected some time well after the release of the gas (i.e., after an equilibrium state of dispersion has been achieved), depending upon the physical characteristics of the chemical and the room, it may not be uniformly spread throughout the room. Thus, a sample taken at any single randomly selected location will not give a representative result for the room as a whole, or even necessarily a good approximation.

1.8.2. Only when the chemical is uniformly dispersed throughout the three dimensions of the room, and is held static in that condition, can a representative result be arrived at from a single sample. The analytical error or measurement uncertainty would also need to be negligible when analyzing the one sample. In all other cases, the true population mean (μ)[*]

[*] Appendices C and D.

*(*the real average concentration for the room as a whole) must be approximated by averaging the results from a number of samples.

1.8.3. The greater the variability in the chemical concentration throughout the room is, the more individual samples will be required to formulate an accurate approximation of the true average. Therefore, as decision confidence requirements increase (i.e., as confidence increases toward 1 or 0 decision error tolerance), the number of samples required to correctly estimate any statistical parameter will also increase.

1.8.4. Variability is a measure of the degree of dispersion (or spread) for a set of values. The sample variance[*], s^2 and sample standard deviation, s, measure the spread of individual measurements or values about the sample mean[†], $\bar{x}$. Some factors that may contribute to variability in environmental populations are the following.

1.8.4.1. Distance, direction, and elevation relative to point, area, or mobile population sources.

1.8.4.2. Non-uniform distribution of pollution in environmental media owing to topography, hydrogeology, meteorology, actions of tides, and biological, chemical, and physical redistribution mechanisms.

1.8.4.3. Diversity in species composition, sex, mobility, and preferred habitats of biota.

1.8.4.4. Variation in natural background levels over time and space.

1.8.4.5. Variable source emissions, flow rates, and dispersion parameters over time.

1.8.4.6. Accumulation or degradation of pollutants over time.

1.8.5. For a particular sampling plan where n measurements are taken for some contaminant of concern in a study area, a (sample) mean concentration ($\bar{x}$) and (sample) standard deviation (s) for the contaminant are calculated. The standard deviation measures the variability of the individual measurements. However, it is often the case that it is the variability of $\bar{x}$ itself that is of interest. The variability of the mean is often measured by the standard deviation of the sample mean, $s_{\bar{x}} = s/\sqrt{n}$. Those two sample values, $\bar{x}$ and $s_{\bar{x}}$, are used to estimate the interval (range) within which the true mean (μ) of the chemical concentration probably occurs, under the assumption that the individual concentrations exhibit a normal (bell-shaped) distribution.

[*] Appendices D, E, and H.
[†] Appendices C, D, E, F, G, and H.

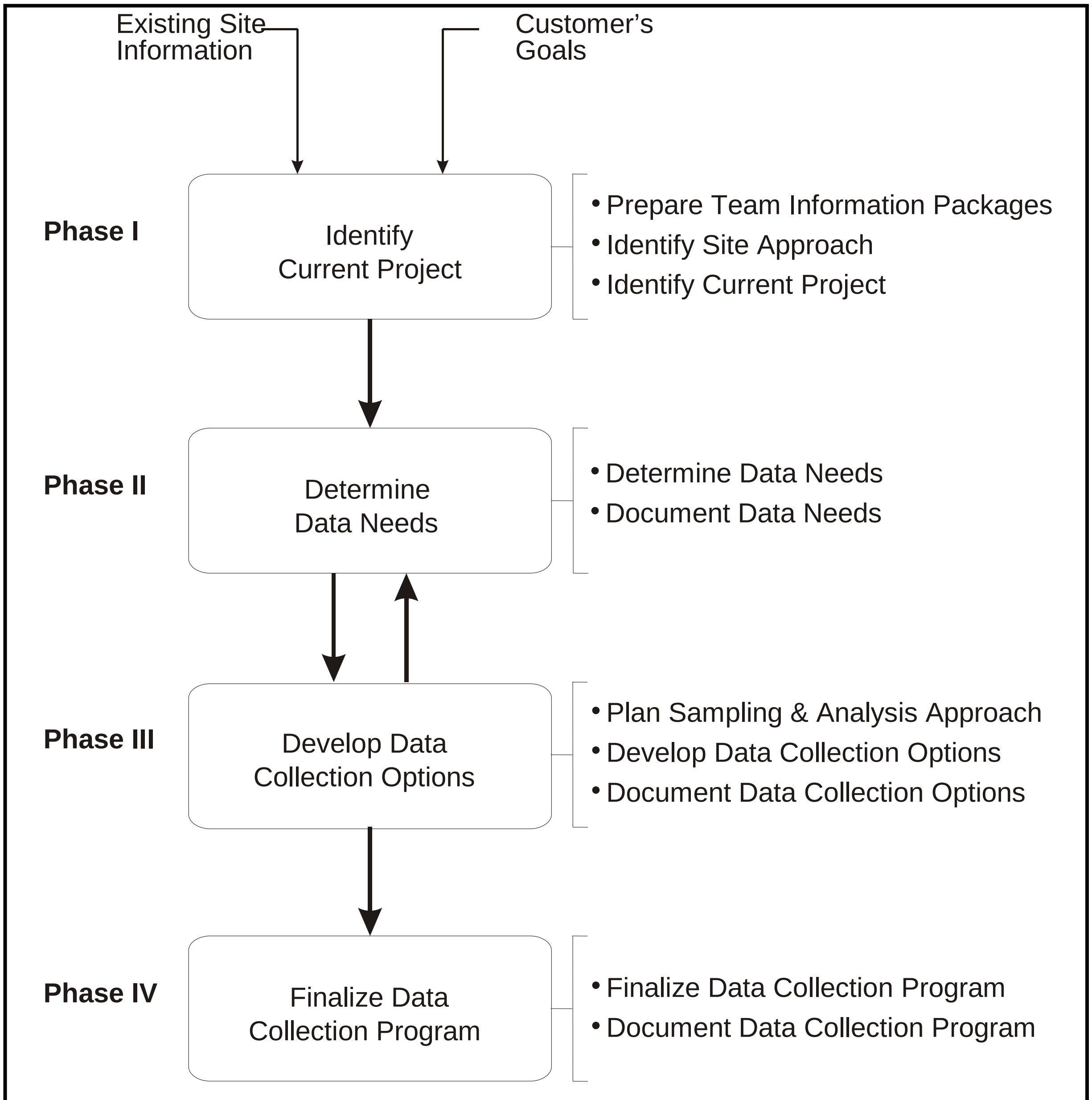
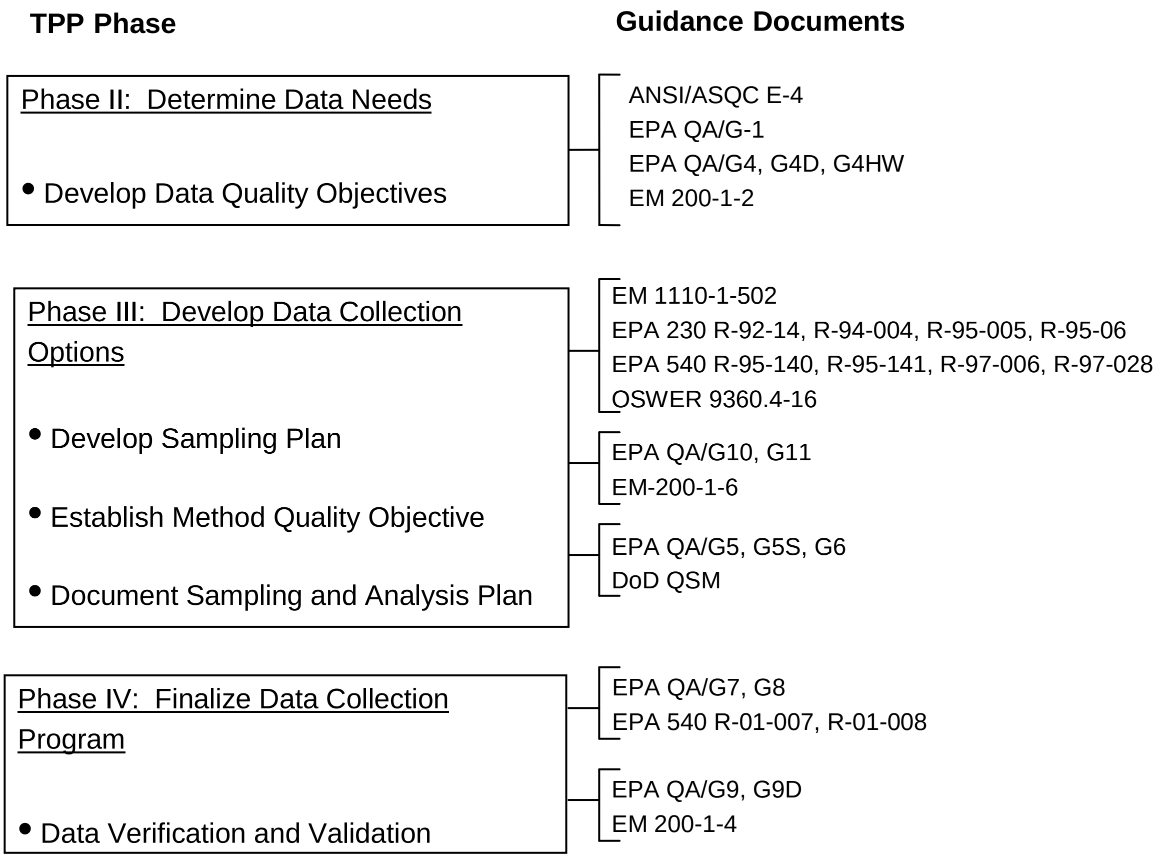

Existing Site Information
Customer's Goals
Phase I
Identify Current Project
• Prepare Team Information Packages
• Identify Site Approach
• Identify Current Project
Phase II
Determine Data Needs
• Determine Data Needs
• Document Data Needs
Phase III
Develop Data Collection Options
• Plan Sampling & Analysis Approach
• Develop Data Collection Options
• Document Data Collection Options
Phase IV
Finalize Data Collection Program
• Finalize Data Collection Program
• Document Data Collection Program

Figure 1-1. Project Planning Life

1.8.6. The relationship among variability, available resources (expressed as sample number, *n*), and decision confidence or lack of uncertainty is fundamental to the project planning process. In general, cost increases as the desired level of confidence or lack of uncertainty increases. Thus, balancing cost and confidence is a primary objective of the planning process. As illustrated in Figure 1-3, this can be depicted as a balance between cost and level of uncertainty: reducing uncertainty increases project costs. As the number of samples increases, the uncertainty decreases but the cost increases. As depicted in Figure 1-3, project planning is the fulcrum of a seesaw balancing cost and uncertainty.



TPP Phase

Guidance Documents

Phase II: Determine Data Needs

- Develop Data Quality Objectives

ANSI/ASQC E-4
EPA QA/G-1
EPA QA/G4, G4D, G4HW
EM 200-1-2

Phase III: Develop Data Collection Options

- Develop Sampling Plan
- Establish Method Quality Objective
- Document Sampling and Analysis Plan

EM 1110-1-502
EPA 230 R-92-14, R-94-004, R-95-005, R-95-06
EPA 540 R-95-140, R-95-141, R-97-006, R-97-028
OSWER 9360.4-16

EPA QA/G10, G11
EM-200-1-6

EPA QA/G5, G5S, G6
DoD QSM

Phase IV: Finalize Data Collection Program

- Data Verification and Validation

EPA QA/G7, G8
EPA 540 R-01-007, R-01-008

EPA QA/G9, G9D
EM 200-1-4

Figure 1-2. Guidance Document Life Cycle.

1.8.7. When dealing with regulators and clients, it is often beneficial to illustrate, in mathematical terms, the relationship among the project objectives, the desired confidence for decisions, and the cost of the project.

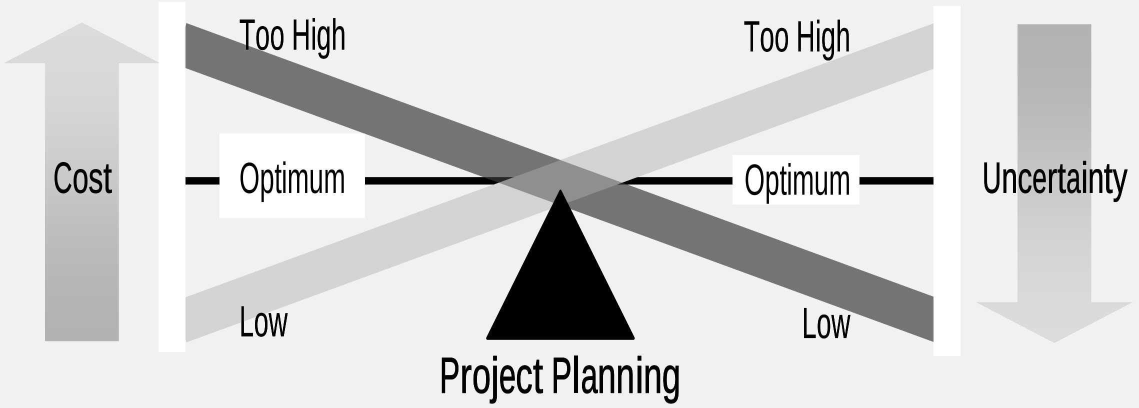


Figure 1-3. Balance Between Resources and Certainty

1.8.8. Figure 1-4 illustrates the relationship of factors that need to be considered in successful project planning.

1.8.9. The purpose of the project planning triad approach is managing total decision uncertainty. Total uncertainty may be viewed as the sum of analytical and field uncertainty. Analytical uncertainty is the portion that arises from variability and bias in the instrumental or analytical test method (as indicated in Table 1-3). Field uncertainty depends on factors such as the temporal and spatial variability of the target environmental population (Table 1-3). Field variability typically exceeds the analytical variability and primarily depends on the sampling design (e.g., the total number of samples, the sample mass, and the nature of field sampling and laboratory sub-sampling methods). In general, data produced by screening analytical methods will contain more analytical variability and bias than data produced by definitive methods. However, field analyses are less costly than laboratory analyses, so a greater number of field samples can be analyzed than laboratory samples for the same fixed cost. Thus, even though field analyses typically contain higher analytical variability relative to laboratory analyses, a larger number of field samples can reduce the total variability more effectively than a smaller number of similarly collected laboratory samples. Field analytical methods should be scrutinized, however, because the total uncertainty does not depend on measurement precision (variability) alone; it also depends on a number of data quality elements such as analytical bias, sensitivity, and specificity (i.e., the ability to detect or quantify the analyte or contaminant of concern in the presence of other analytes or interferences in the sample).

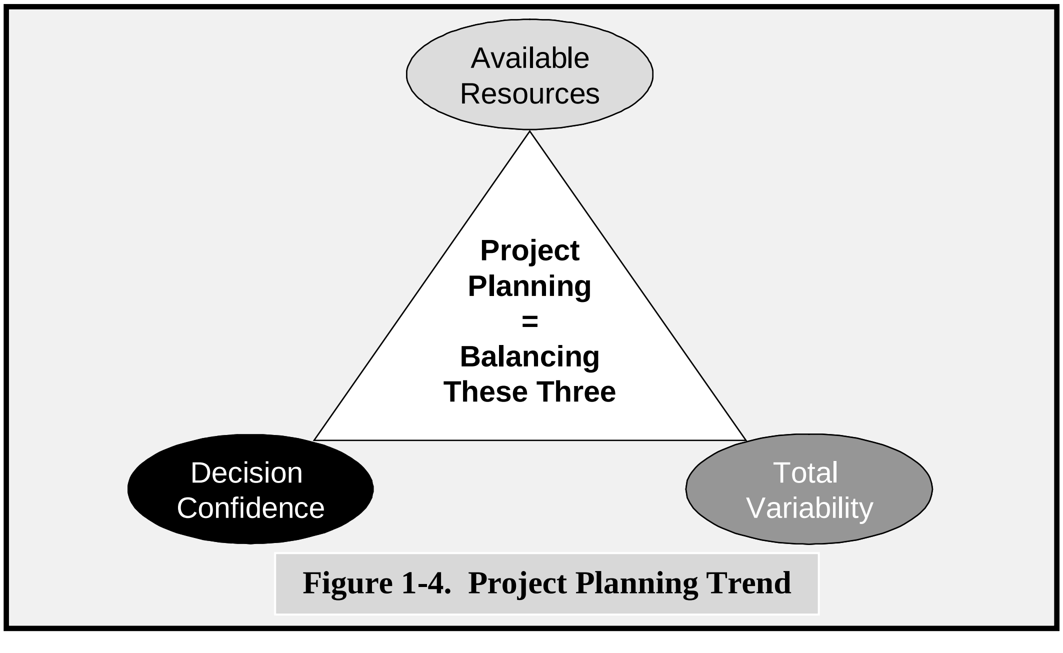


Figure 1-4. Project Planning Trend

1.8.10. The triad approach also makes use of rapid turn-around times for field methods. Field methods have an advantage over laboratory methods in that they are capable of providing data to support decisions while mobilized in the field. For example, managers can modify sample locations on the basis of new information about the extent of contamination during a single mobilization. In contrast, fixed laboratory data packages are produced several weeks after sampling is complete. Remobilization may be necessary to resolve questions arising from laboratory results.

1.8.11. The triad approach is especially useful for statistical designs such as adaptive sampling,* ranked set sampling*, and systematic sampling*, as these designs often require larger numbers of samples. To successfully implement the approach, the capability of the field methods must be scrutinized with respect to project data quality and measurement objectives. For example, many field methods are not as sensitive or selective as laboratory methods. If the primary objective is to characterize contamination with respect to some fixed risk-based limit or cleanup goal, and the detection limit is greater than the decision limit, then comparisons of the field data to the decision limit will not be viable. Comparisons of field and laboratory data during a pilot test phase to verify or establish correlation between two sets of results is a useful approach for evaluating and selecting field methodologies.

1.8.12. The triad approach relies on thorough, systematic planning to articulate clear project goals and encourages negotiations among stakeholders to determine the desired decision confidence. A multidisciplinary technical team then determines what information is needed to meet those goals. A key feature of this planning is identifying what uncertainties could compromise decision confidence and allowing team members with appropriate sampling and analysis expertise to explore cost-effective strategies to minimize them. Often, the most cost-effective work strategy involves the second leg of the triad, which is using a dynamic work plan to make real-time decisions in the field. The third leg of the triad uses field analytical methods to generate real-time on-site measurements that support the dynamic work plan. Projects managed using these concepts have demonstrated cost savings of up to 50% over traditional approaches.

1.8.13. The contributions to the total variability (i.e., the total precision component of the uncertainty) can be expressed as a vector sum of an analytical component and sampling component of the variability (e.g., or as a ratio of the sampling to analytical variability, say 9:1). Although the analytical variability is minimized by conventional laboratory analyses, sampling variability is often not adequately addressed. Budget constraints invariably limit the number of laboratory analyses. A combination of high laboratory analysis costs and a poor sampling design often results in a low sampling density that is not very representative of the environmental population of interest. Field studies consistently find that the sampling design, rather than analytical considerations, predominately governs the total variability.

1.8.14. When analytical costs are lower, more samples can be analyzed, yielding more confidence in the representativeness of the data set (Phase 1). This is most effective if field methods are used to generate data and a dynamic work plan rapidly resolves any uncertainty about location and volume of contamination (for example, locate and delineate hot-spots in a

* Appendices C and D.

single field mobilization). If the analytical data quality used to manage sampling uncertainty is less than what is eventually needed to make final project decisions, such as whether the site can be declared clean, more expensive definitive analyses may be performed on samples selected to refine the feature of interest (Phase 2). However, if the initial method produces data of sufficient rigor to support defensible decision-making, then additional, expensive analyses would be redundant and unnecessary.

1.8.15. In Phase 1, analytical uncertainty (variability) increases so that unit sample costs decrease, allowing a higher sampling density than with the conventional approach. As a result, sampling uncertainty (variability) decreases, lowering the overall uncertainty in data interpretation. Sampling uncertainty is further decreased if hot-spot removal reduces the variability in contaminant concentration and if representative sampling locations for more rigorous analysis are identified based on Phase 1 information. The vector representation of uncertainty for this approach indicates that the overall uncertainty in the data set for site decision-making will be much less than the overall uncertainty in the conventional method.

1.8.16. Data quality should be judged on whether both the sampling and the analytical uncertainties in the data sets support decision-making at the desired degree of decision confidence. However, relying solely on regulator-approved, definitive analytical methods, while ignoring sampling uncertainty, easily produces uncertain decisions.

1.8.17. When field analytical methods are used, the process and resulting data are often referred to as “field screening.” The term is misleading when field methods are of adequate quality to satisfy project DQOs; field analyses are not necessarily “screening” or inferior to fixed-laboratory analyses in the context of the overall end use of the data. Here, alternate terminology is proposed to reflect current EPA guidance that both sampling and analytical uncertainties must be managed to assess data quality. We consider the two terms “effective data” and “decision-quality data,” to be equivalent when describing data of known quality that are effective for making defensible primary project decisions, because both sampling and analytical uncertainties have been explicitly managed to the degree necessary to meet clearly defined project goals.

1.8.18. Primary project decisions are those decisions that drive resolution of the project, such as whether or not a site is contaminated and what subsequent actions, if any, will be taken. Therefore, contaminant data are usually the data sets of interest. But data sets can interact in complex ways, and are referred to as collaborative data sets. For example, a contaminant data set considered alone might not be effective for making project decisions, yet the same data set might be more effective when combined with other data or information to manage the remaining uncertainties. Ancillary data refers to data used to support many other project decisions that fall under worker health and safety monitoring, data that help in the understanding of fate and disposition of contaminants, and data that aid in decisions about the representativeness of environmental samples.

1.8.19. This decision-making paradigm and terminology embodies the central theme of systematic project planning, the management of decision uncertainty.

CHAPTER 2

Preliminary Assessment and Site Investigation

Section I
Preliminary Assessment

2.1. Introduction. A Preliminary Assessment (PA) is initiated after a CERCLA site (or suspected site) is identified. Statistical evaluations are not typically conducted for a PA. The purpose of the PA is to determine if a site poses a potential threat to human health or the environment. EPA maintains a list of actual and potential hazardous substance releases requiring CERCLA response. The property owner or agent is obliged to perform a PA; for Federal facilities, a PA is required within 18 months of listing (57 FR 31758; 17 July 1992).

2.1.1. The PA process collects information from existing resources. Generally, PA data are qualitative rather than quantitative, and do not require statistical evaluation. In some instances, historical chemical data may be available, but the PA does not require that such data be statistically manipulated. The EPA evaluates the site information according to the Hazard Ranking System (HRS) as detailed in 55 FR 51531 (14 December 1990). HRS calculations do not have statistical components. Some examples of PA information necessary to the HRS are as follows.

2.1.1.1. Identification of wastes or waste sources.

2.1.1.2. Physical site conditions, such as precipitation rates, depth to groundwater, or distance to surface water bodies.

2.1.1.3. Workers or residents at a site.

2.1.1.4. Local population within a set radius of a site.

2.1.2. Based on the results of the HRS, a site may warrant further investigation or no further action. Though quantitative statistical evaluations are not required during a PA, the following case study illustrates the value of a thorough qualitative evaluation of PA information.

2.2. Case Study 1—Examining Historical Data Sets. In the preliminary assessment of a landfill located on a manufacturing facility in Pennsylvania, some historical analytical data were available to the project team. The question raised, however, was whether or not those data would be usable in the PA. If the data were found to be usable and applicable, the landfill might be removed from further consideration in the CERCLA process. However, if the data were not found to be usable, then a Site Inspection (see Section II) would be needed.

Moreover, if the data were used, prior to further validity testing (thus, explicitly assuming the data were reliable), and found later in the assessment to be erroneous, inaccurate and misleading conclusions would have been drawn.

2.2.1. Several different assessments of the data were required: i) Were the precision, accuracy, and representativeness of the data sufficient for the purpose? ii) Was the sampling design for the historical data sufficient for the purpose? and iii) Were the data comparable from historical event to historical event and could they be combined with new data, if necessary, to draw conclusions about the site?

2.2.2. The existing data were included in monitoring reports to the state. The reports consisted of little more than sample identification, date, and analytical results. Only positive detections were reported. Based on that information alone, the project team could not assess the quality of the data and concluded that unless additional information was obtained, the data could not be used as part of the PA. The site owners began to investigate the origins of the data.

2.2.3. In the interim, the project team assigned a geologist to examine the sampling design for the work. The facility had identified a single monitoring well, MW-02, as an upgradient location for comparison to a set of three downgradient wells, MW-03, MW-06, and MW-08. Through a review of well construction diagrams, as well as available topographic and hydrogeologic information, the geologist found that the well identified as upgradient was located within 3 feet of the landfill footprint, in a swale that received run-off from the landfill.

2.2.4. Thus, it was likely that the upgradient well was directly impacted by landfill operations and would not constitute an acceptable upgradient location. Further, MW-06 and MW-08 were found to be generally cross-gradient to MW-02 rather than directly downgradient, and that MW-03 had been screened in a perched aquifer, hydrologically isolated from the aquifer monitored by the other three wells.

2.2.5. Upon receipt of laboratory data packages for the historical data, the project team observed that a variety of different analytical methods and laboratories had been employed in the course of the work, resulting in mixed reporting limits and inconsistent detection of analytes. As a result of these assessments, the historical data were judged not to be usable for the PA.

2.2.6. In summary, prior monitoring appeared to indicate the presence of contamination (e.g., which would have triggered an RI), but additional evaluation data indicated that the data were not usable; therefore, an SI was initiated.

Section II
Site Inspection

2.3. Introduction. The Site Inspection (SI) is the next step in the CERCLA process. Statistical evaluations are often appropriate for an SI. Typically, the major objective of these evaluations is to establish the presence or absence of site contamination with respect to predefined decision limits. An SI is performed if the PA indicates the potential for hazardous materials to be present, if human or ecological receptors, or both, exist, and if there are potential complete exposure pathways for the receptors. The SI generally focuses on establishing, through sampling and analysis, whether hazardous materials are present at concentrations that exceed some “screening criteria.” The project planning team must establish decision limits or screening criteria prior to sampling and analyses. Generally, decision limits fall into the following categories:

2.3.1. Naturally occurring or known background levels (site-specific background information is typically unavailable at the SI stage).

2.3.2. Ecological benchmarks, which are dependent on analytes and media (typically developed with regulatory input).

2.3.3. Risk-based screening criteria for human health such as EPA Region IX Preliminary Remediation Goals (PRGs) or EPA Region III Risk-based Concentrations (RBCs) are available at the following Web sites.

http://www.epa.gov/region09/waste/sfund/prg/index.html

http://www.epa.gov/reg3hwmd/risk/index.htm

2.3.4. Applicable or relevant and appropriate requirements (ARARs). For example, Maximum Contaminant Levels (MCLs) for drinking water may be ARARs for some CERCLA sites.

2.3.5. During the DQO process, stakeholders identify the study questions, such as the presence or absence of contamination with respect to a set of decision limits, the nature and quantity of the data required to support the decision-making process, and the acceptable tolerances for decision errors. Selecting the screening criteria is critical for establishing both data quality objectives (DQO) and measurement quality objectives (MQOs). MQOs are established after DQO development. MQOs for analytical sensitivity must be adequate to report quantitative contaminant concentrations at levels less than the project decision limits. (Refer to Appendix G for a discussion of detection limits and quantitation limits.)

2.3.6. Team members must establish the DQOs for the project at the outset of the SI. In an SI, stakeholders must identify the problem at the site and how it will be evaluated, identify the decisions to be made using the data, and specify limits on that decision error. These will lead the project team to an optimal sampling design at a site. Appendix G discusses detection limits, quantitation limits, and censored data. Understanding the concepts in the context of ARARs guides part of the project planning.

2.4. Sampling Design. In general, statistical sampling designs are required to support statistical evaluations. Professional judgment, site-specific information, and DQOs must be used to select the type of the statistical sampling design (e.g., random* as opposed to systematic sampling) and the required number of samples. The sampling design depends on factors such as the nature and distribution of the contamination in the study area, sampling cost, tolerances for decision error, and perceived level of decision uncertainty. For example, a small number of samples during the SI stage may be beneficial for short term cost considerations, but may not be adequate to achieve the desired tolerances for decision uncertainty and error and may, therefore, not be a cost-effective strategy by project closeout (as multiple sampling events rather than a single sampling event would typically be required to support decision-making).

2.4.1. Decision uncertainty refers to statistical variability, subjective judgment, randomness in the process, disagreement, and even imprecise wording inherent in the decision-making process (Moser 2000). Decision uncertainty is a function of the variability of the contaminant of concern in a study area and depends on the number of samples collected. For example, if the sample mean, $\bar{x}$, is an appropriate measure of site-wide contamination and the standard deviation of the sample mean, $s_{\bar{x}}$, measures the variability around $\bar{x}$, then the variability (and uncertainty) decreases as the number of samples *n* increases, because $s_{\bar{x}} = s/\sqrt{n}$. (Increasing the physical size of each sample would also decrease the variability.) It should also be noted that, in addition to decreasing the variability, $\bar{x}$ becomes a more accurate estimate of the population mean, μ, as *n* increases.

2.4.2. Site-specific information must be taken into account when selecting the sampling design. In particular, the team members need to identify potential source areas and any stratification they may represent. For example, suppose there are two sources of lead at a bomb reconditioning facility—stack emissions affecting surface soil and old buried waste piles affecting subsurface soil. This information can be used to design a sampling scheme for the "surface soil stratum" and a separate scheme for the "subsurface soil stratum." Likewise, there may be different study objectives for each stratum. Surface lead may be of concern for exposure of site workers and subsurface lead may be of concern for protection of groundwater. Stakeholders would need to identify these issues during project planning to develop an optimal site-wide sampling design.

2.4.3. Several different types of sampling designs are listed below. Appendix C presents a detailed explanation of these designs.

a. Judgmental sampling.

b. Random sampling.

* Appendices C and D.

c. Simple random sampling.

d. Stratified random sampling.

e. Systematic and grid sampling.

f. Ranked set sampling.

g. Adaptive cluster sampling.

h. Composite sampling.

2.4.4. The TPP and DQO processes are used to develop an appropriate sampling design for the SI phase. Two case studies are presented below to illustrate sampling designs commonly used for SI.

2.5. Case Study 2—Judgmental Sampling, Oil/Water Separator. Project planners found an oil/water separator buried underground at a pipe mill. There was evidence of leakage to the surface soils around the tank and a release to groundwater was suspected. The objective was to determine if there was a measurable presence of oil floating on the water table.

2.5.1. Historical information and local knowledge allowed a hydrogeologist to determine the direction of groundwater flow. The hydrogeologist also knew of two monitoring wells in the area. One well was located upgradient to the separator; the second was cross-gradient.

2.5.2. The project planners decided to place a new monitoring well downgradient of the separator. Because they were looking for an oil product, the soil boring for the monitoring well was logged by a geologist who could then identify the water table depth. The well was installed so that the screen intersected the water table, where floating oil would most likely be visually detected.

2.5.3. Judgmental sampling was predominantly used in this example because the planners possessed significant existing site information. They knew the physical properties of the oil, they knew the hydrogeology of the site, and they were answering a no quantitative question.

2.5.1. Case Study 4 predominantly illustrates the application of composite sampling[*] and stratification[†] for a SI, and the iterative nature of the DQO process when optimizing a sampling design.

2.6. Case Study 3—Arsenic Contamination in Soil. At an active manufacturing site, arsenic contamination was widespread in surface soils. Preliminary screening analyses and risk

[*] Appendices C and D.

[†] Appendix D.

assessments identified worker exposure as the most likely concern. The site was initially divided (stratified) into 90 subunits related to work areas for a more in-depth evaluation of risk. Based on financial constraints, the project team was allocated a budget of $50,000 for SI sampling and analytical testing.

2.6.1. The aggregate initial cost of a field grab sample was $175, with $100 attributed to field collection and $75 attributed to laboratory analysis. The expected percent relative standard deviation (%RSD) for the analytical (laboratory) measurements was 5%. The estimated standard deviation, *s*, for the analytical method, at the decision limit of 600 ppm, was computed as 5% of 600 ppm or 30 ppm.

2.6.2. The planning team estimated the field component of the variability to be 10 times greater than the laboratory component of the variability. Thus, the %RSD for the field component of the variability was calculated by multiplying the %RSD for the analytical measurements by 10 (yielding a field component %RSD of 50%). This estimate was then multiplied by 600 ppm to yield a value of *s* equal to 300 ppm for the field component of variability (i.e., 50% of 600 ppm). The estimates for field and analytical variability (i.e., variance or s^2) were then combined and the standard deviation was calculated (s = 330 ppm). The maximum observed arsenic concentration was 720 ppm. The analytical method was deemed appropriate by the planning team. If historical sampling data were available, the data would be used to estimate the field variance and to test for normality.

2.6.3. The planning team principally considered two sampling design alternatives—simple random sampling and composite sampling (see Appendix C for a review of each sampling method). A *t*-test was used to calculate the sample size for simple random sampling (Appendix L). Given a decision error limit of α = 0.01, more than 200 samples per work area would have been required (refer to Appendix L for a review of methods involved in setting and testing hypotheses). The total cost of this sampling effort would have exceeded $3 million.

2.6.4. Using similar methods, the team explored composite sampling, which would have required 30 samples to be collected per work area for a cost of over $1 million. Given the considerable cost burdens for both proposed sampling designs, the team decided to return to Step 6 of the DQO process and modify the decision error limits. The team found that by increasing α to 0.05, the composite sampling design would require the collection of 13 samples for each of the 90 work areas. This revised design had a total cost of $204,750, approximately one-fifth of the original estimate.

2.6.5. The team realized that they would have to find other means of generating an appropriate design while remaining within budget. To do this, the project team redefined the boundaries of the study (by revisiting Step 4 of the DQO process). The team recognized that one of the drivers of the cost was the large number of separate study units (previously, the calculated sample size was applied to each of the study units). The planning team used exposure information for the contaminant to map out the potential or expected pathways in the surface soils through which the contaminant could spread. The potential pathways were categorized into four distinct spatial units.

2.6.6. Rather than collect data and make decisions for each of the 90 individual work areas, the team decided to sample and make decisions for each of the four risk areas. Recognizing that these larger areas carried greater decision error consequences, the team revisited Step 6 of the DQO process and established new limits for decision errors applicable to the four risk areas. The team established different decision confidence limits for each and recalculated the number of samples required. The cost of implementing this design was $38,850, which fell within the $50,000 budget for the sampling and analysis.

2.7. General Review of Sample Size Determination[*]. For typical statistical sampling designs, there are well-defined relationships between the number of required samples (i.e., sample size), tolerance for decision errors, and inherent variability of the analytical measurements and the target environmental population. One such relationship states that the sample size increases as the tolerance for decision error decreases or the variability increases. The sample size must be equal to or greater than the sample size required to achieve predetermined tolerances for decision errors. When confidence limits for the mean are of interest, an appropriate sample size is required to generate a sufficiently precise estimate of the true mean concentration of a chemical contaminant (refer to Paragraph 3.11 and Appendix K for additional discussion of confidence limits). For the example presented above, the sample size must be adequate to demonstrate that the upper limit of the CI for μ is less than the applicable regulatory threshold, RT. The required sample size must increase as s^2 increases and as the difference Δ (RT – $\bar{x}$) decreases. In a well-conceived sampling plan for a solid waste, every effort should be made to estimate the values of $\bar{x}$ and s^2 before sampling starts. Case Study 3 illustrated that decision confidence affects sample size. Case Study 4 illustrates this concept in a different setting.

2.8. Case Study 4—Effect of Decision Confidence on Sample Number. Upon promulgation of the Toxicity Characteristic Leaching Procedure (TCLP) rule, a steel mill in Maryland contracted with a consultant to collect samples from various waste streams within the facility for TCLP analysis of metals (this case study considers only the cadmium data). One such waste stream was from a wastewater treatment system and consisted of collected sludges. Although no previous analysis of sludges had been done, cadmium had been monitored in the wastewater stream before treatment. The project manager believed that the wastewater data would be sufficient for establishing routine variability of cadmium in the sludge, assuming there were no great differences in the treatment process over time and a 10 times concentration factor from wastewater to sludge.

2.8.1. The project manager decided to use the past year's wastewater data to make

[*] Appendix L.

preliminary estimates of the number of samples required to meet the statistical confidence requirements of the TCLP rule (i.e., $\alpha = 0.2$). Four results (in milligrams per liter [mg/L]) were available from the previous year as follows: 14.2, 9.6, 21.7, and 19.3.

2.8.2. The mean and variance of the results (as adjusted for concentration to sludge) were the following: $\bar{x} = 1.6$ mg/L and $s^2 = 2.2$ mg/L, respectively. The proposed water regulatory threshold value (RT) was 1 mg/L. Using the formula for simple random sampling, the project manager calculated the number of samples required as follows:

$$n = (t^2 \times s^2) \div (\mathrm{RT} - \bar{x})^2$$

where: n = number of samples required
t = Student's value for n–1 degrees of freedom and 0.8 confidence
s^2 = sample variance
$\bar{x}$ = sample mean
RT = regulatory threshold.

2.8.3. Thus, $n = [(0.9785)^2 \times 2.2]/(1 - 1.6)^2 = 6$ samples. Samples are an integer value, and should be reported without decimal fractions. (The value of t may be obtained from Table B-23, where $df = 3$ and $p = 0.8$.) Assuming a sampling cost of \$50 per sample and an analytical cost of \$25 per sample, this testing would cost \$450.

2.8.4. The client's attorneys asked what the effect would be should they wish to establish a safety margin by increasing the decision confidence to $\alpha = 0.05$. The revised plan would require

$n = [(2.353)^2 \times 2.2]/(1 - 1.6)^2 = 34$ samples, or a sampling and analysis cost of \$2,550.

2.9. <u>Summary of Case Studies</u>. Case studies 2 through 4 illustrate the multitude of related factors that must be considered when evaluating which sampling design to apply in a particular SI. When evaluating alternative sampling plans, planners may anticipate the concentration patterns likely to be present in the target population. Advanced information about these patterns can be used to design a plan that will estimate population parameters with greater accuracy and less cost than can otherwise be achieved.

2.10. <u>Comparing On-site Data to Fixed Screening Criteria</u>. In the data analysis phase of the SI, environmental scientists compare site data to screening values using either qualitative or quantitative statistical evaluations. The following provides a discussion of qualitative and quantitative evaluations.

2.10.1. <u>Qualitative Statistical Evaluations</u>. The EPA has developed risk-based screening criteria in the form of PRGs and RBCs. These criteria are frequently applied at the SI stage to identify whether the site as a whole may need further attention in an RI/FS. Many screening criteria exist at both the Federal and state government level. Thus, comparisons are frequently made against the lowest of several screening criteria that can be applied to a given data set from a given location. The technical team must ensure that the criteria are being applied properly (i.e., not all screening criteria are applicable to every site), and that the implications are clear in the conclusions of the SI. For example, if site data exceed a standard developed to protect groundwater from soil leaching of contamination, but do not exceed an applicable human health standard, the team should report the results with the implications of these differences noted in the conclusions.

2.10.2. One typical qualitative method of comparing data decision limits entails the use of a spreadsheet or database. The decision limits and individual sample results are presented in a tabular format and each detected analyte concentration is compared to the corresponding screening values for that analyte. (It may be necessary to compare a single contaminant of concern to only the lowest decision limit or several different decision limits.) Table 2-1 is an example of such a spreadsheet.

2.10.3. The primary pitfall of this qualitative strategy is that the uncertainty associated with the reported results is not considered when the results are compared to the decision limits. Thus, the reported results may actually be equal to or exceed decision limits when uncertainty is taken into consideration. If this is the case, especially in the event the decision limit is exceeded, the wrong conclusion would be drawn. The ramification of an erroneous conclusion will vary, depending on the nature of the problem under investigation; nevertheless, this is an outcome that should be avoided or at least minimized.

2.10.4. Historically, environmental researchers have tended to screen analytical results into two categories—greater than the standard or less than the standard. Through advances in research and technology, three categories now exist against which analytical results can be compared: i) the reported value clearly exceeds the standard (when bias and variability are taken into account); ii) the reported value clearly does not exceed the standard; and iii) the result is inconclusive. This last conclusion is reached when the uncertainty is too large for reliable decision-making.

2.10.5. Table 2-1 illustrates how qualitative information may be used to support the decision making process when SI data are qualitatively, rather than statistically, compared to decision limits. In particular, information regarding the quality of the data, obtained in the data validation process, is used to determine whether contamination is present at concentrations greater or less than project decision limits. All applicable screening criteria are displayed in Table 2-1. For example, the “S” column reports the results of comparing each analyte concentration and the lowest screening limit. One of three codes is entered in this column for the three possible conditions identified in the preceding paragraph. An “X” is recorded if the reported values appear to be well above the decision limit, an “I” if the result is inconclusive, and a blank space if the result appears to be well below the limit. Select results from Table

2-1 are discussed below to illustrate the nature of the screening evaluation.

2.10.5.1. Tetrachloroethane results in IRP-49 (1.2 ppb) and IRP-51 (17.08 ppb) both exceed the PRG (1.1 ppb). Although the value in IRP-49 is barely above the PRG, it reports the results as two significant figures, so we must accept its value as exceeding the PRG. However, accounting for analytical error, typically between 20 and 30% (as a conservative estimate), this result would be inconclusive. The researcher then must choose whether to conduct additional testing or accept the value of IRP-49 as an exceedance. The latter would be selected only if a conservative estimate was desired.

2.10.5.2. In IRP-49 (0.2 ppb) and IRP-51 (0.2 ppb), the reported concentration is not distinguishable from the PRG when compared on the basis of just one significant figure. Therefore, these results are inconclusive.

2.10.5.3. Several chloromethane results are marked inconclusive because of blank contamination. The only sample without blank contamination, IRP-39, was below the PRG (PRG = 1.5 ppb; IRP-39 = 0.2 ppb). The reported concentration was qualified with a J flag because it is less than the quantitation limit of 1 ppb. (The quantitation limits are not listed in Table 2-1, but were obtained from the laboratory’s data package.)

2.10.5.4. For bromodichloromethane in sample IRP-48 (0.2 ppb), the reported concentration is biased low and is less than the quantitation limit of 1 ppb, so this exceedance of a PRG (0.18 ppb) is conclusive. In sample IRP-51 (0.1 ppb), the result is also biased low and is just below the PRG, so this result is also not conclusive.

2.10.5.5. For chloroform in sample IRP-39 (0.4 ppb), the reported concentration is qualified with a J flag because it is less than the quantitation limit of 1 ppb. As the reported result is quantitatively estimated, it does not reliably demonstrate that chloroform is present above the PRG.

2.10.5.6. Benzo(a)pyrene was reported in sample IRP-49 (0.278 ppb) above the PRG limit (0.0092 ppb). However, the detection limit (0.014 ppb) is above the PRG for the remaining samples. Only by achieving a lower detection limit is it possible to determine whether the non-detects are a problem. The results for benzo(a)pyrene are marked inconclusive. All of the arsenic non-detects are inconclusive based on a similar rationale.

2.10.5.7. Though the reported concentration of chloride in sample IRP-49 (265 mg/L) is not qualified as estimated and exceeds the decision limit (250 mg/L), the result is marked inconclusive because the difference between the detected concentration and the decision limit is less than 5%, which is smaller than the analytical error for the test method (e.g., the error tolerance for the test method is typically 5 to 20%).

2.10.6. These results illustrate the critical importance of estimating and incorporating into decision-making knowledge of both the field and laboratory components of variance.

One fundamental error is treating the reported results as conclusive when in fact they are not. The values represented in this table are measurements, and measurements contain bias and variability that must be accounted for in decision-making. (See EM 200-1-10 for additional guidance on the data review strategies that were primarily used to qualify the results in Table 2-1.)

2.11. Quantitative Statistical Evaluations. When the results of the qualitative statistical evaluations are inconclusive, further investigation is required. DQOs must be revised so that the parameter of interest is no longer a single datum per location. Instead, multiple samples are collected for those uncertain locations and the resulting distribution of values is compared to the decision limit using quantitative statistical tests. The results would typically be statistically compared to decision limits using one-sample tests[*] for central tendency, as discussed below.

2.11.1. All statistical tests require the user to make certain assumptions about the data to perform the statistical test. The user must demonstrate that the underlying assumptions for a particular statistical test are reasonable before doing the test. With respect to these underlying assumptions, statistical tests can be roughly categorized as either parametric or non-parametric.[†] When non-parametric tests are conducted, data sets are required to satisfy fewer assumptions than for the corresponding parametric tests.[†] In particular, a parametric statistical test assumes a specific distribution[†] for the data (i.e., the entire population is described by some specific mathematical function), such as the bell-shaped curve for the normal distribution[‡]. Statistical plots of actual measured sample concentrations must be substantively consistent with the corresponding plots generated using the theoretical functional relationship. Tests that require normal or log normal distributions are most commonly used. (A data set is log normal if, when the log of each datum is calculated, the resulting set of values is normally distributed.) Common graphical methods (i.e., plots) are presented in Appendix J. In addition, an overview of the evaluation of distribution assumptions is presented in Section III of Chapter 3.

2.11.2. It should also be noted that parametric tests become problematic, and may not be possible to perform, when the data sets contain a significant number of censored[§] values (i.e., analyte concentrations reported as non-detects). However, as described in Appendix H, it may be possible to use the Poisson distribution[**] for highly censored data. Parametric tests

[*] Appendix L.

[†] Appendices L and M.

[†] Appendix E

[‡] Appendices E, F, and J.

[§] Appendix H.

[**] Appendice E

are also problematic when there are outliers. The possibility of outliers[*] should be considered ered in every analysis.

2.11.3. Non-parametric tests do not assume a specific functional relationship for the data distribution. These tests tend to be less sensitive to outliers and non-detects than parametric tests. Although non-parametric tests are more applicable relative to parametric tests, non-parametric tests tend to be less statistically powerful than parametric tests. In essence, this means that more samples must be collected for a non-parametric test relative to the corresponding parametric test to make decisions at the same level of confidence.

2.11.4. Background concentrations of naturally occurring and anthropogenically derived compounds are also possible screening criteria. However, there are few instances in which such background levels are available at the SI stage. Sometimes a “site-wide” statistical background study has been done. If such a study is available, two-sample statistical tests[†] would be used to compare the study area data set with the “site-wide” background data set. (As the name implies, a two-sample statistical test is predominantly a statistical evaluation to compare two separate sets of data.) Because an RI often includes specific sampling for background, the determination of background levels and their usefulness is described in Chapter 3. If the SI is the first sampling event for a site, there is a low probability specific background sample data exist.

[*] Appendix I.

[†] Appendix M.

Table 2-1. Site Screening Data Table

		EPA MCL	Region IX PRG (1999)	IRP-39				IRP-48				IRP-49				IRP-51			
Analyte	**Units**	**Tap Water**			**L**	**V**	**S**		**L**	**V**	**S**		**L**	**V**	**S**		**L**	**V**	**S**
Organics																			
Bromodichloromethane	μg/L	—	0.18	0.1	U			0.2	J	L, s	I	0.1	U			0.1	L, s		I
Carbon Tetrachloride	μg/L	5	0.17	0.1	U			0.1	U			0.1				0.4	J	J	I
Chloroform	μg/L	—	0.16	0.4	J	J	I	0.1	U			0.1	U			0.1	U	U	
Chloromethane	μg/L	—	1.5	0.2	J	J		6.1		B	I	1.6		B	I	3.7		B	I
Methylene Chloride	μg/L	5	4.3	0.1	U			0.1	U			0.1	U			0.1	U		
Trichloroethene	μg/L	5	1.6	0.4	J	J		0.1	U			18.7			X	18.1			X
Tetrachloroethene	μg/L	5	1.1	0.1	U			0.1	U			1.2			X	17.1			X
Benzo(a)pyrene	μg/L	0.2	0.0092	0.014	U		I	0.014	U		I	0.278			X	0.014	U		I
Inorganics																			
Arsenic	mg/L	50	0.045	0.7	U		I	0.7	U		I	0.7	U			0.7	U		I
Chloride	mg/L	250	—	311			X	15.8				265			I	134.7			
Lead	mg/L	15	—	0.3	U	K		0.3	U	K		8				10			
Nickel	mg/L	—	730	590				29.0				214				198.0			
Sulfate	mg/L	250	—	44.0				5.98				41.6				21.45			
Thallium	mg/L	2	2.9	1.4				0.8	U			0.8	U			0.8	U		
Vanadium	mg/L	—	260	1.4				1.0	U			3.0				5.0			

Notes: L column contains the laboratory flags. V column contains the validation flags. S column contains screening results.

Flags:
U–Not detected above reported detection limit.
B– Not detected substantially above laboratory or field blank.
L—Biased low.
K—Biased high.
s—Surrogate failure.
J—Quantitatively estimated.

Screening Codes:
X — sample concentration unequivocally exceeds the lowest screening standard.
I — sample concentration comparison to screening standard is inconclusive.
– A blank cell indicates that the sample concentration unequivocally does not exceed the lowest screening standard.

CHAPTER 3

Remedial Investigation/Feasibility Study

3.1. Introduction. If, based on the PA/SI, a site warrants listing on the National Priorities List (NPL), and RI/FS is performed at the site.

3.1.1. The RI is the stage in the CERCLA process for collecting data to do the following.

3.1.1.1. Characterize site conditions (e.g., thickness of unsaturated soil [vadose zone], depth to groundwater, vegetative cover, background conditions).

3.1.1.2. Determine the types, conditions, and distribution of the waste contamination in affected media.

3.1.1.3. Assess risk to human health and the environment.

3.1.1.4. Conduct treatability tests to evaluate the potential performance and cost of the treatment technologies that are under consideration.

3.1.2. The FS is the stage for the development, screening, and detailed evaluation of remedial actions.

3.1.3. The RI and FS are intimately linked. Data from the RI influence the development of remedial alternatives in the FS, which in turn affect the data needs and scope of treatability studies and additional field investigations. This phased approach encourages the planning team to continually plan the site characterization effort, which minimizes the collection of unnecessary data and maximizes data quality.

3.1.4. As in the SI phase, the initial statistical elements in the RI process involve the development of DQOs. The statistical evaluations used for the RI typically include those performed for the SI. For example, as in the SI, site data are often statistically compared to some set of fixed decision limits and upper confidence limits are often established (as discussed in Chapter 2). In general, the statistical evaluations are more common for RIs than SIs, and the statistical analysis tends to be more comprehensive. In part, this is because typically data coverage is greater and the RI data quality objectives are more robust. For example, while the SI predominantly focuses on statistical evaluations to resolve the presence or absence of contamination, the RI reaches for a determination of the extent of contamination. Critical to the onset of an RI is the identification of Applicable or Relevant and Appropriate Requirements (ARARs), which, in turn, may influence the identification of areas requiring remediation. Both sampling strategy and extent of contamination are influenced by the selection of ARARs. ARARs help identify the best analytical procedures needed to reach decision limits. This aspect of DQOs is addressed in Appendix C.

Section I
Site Characterization

3.2. Introduction. The first two objectives of the RI (subparagraphs 3.1.1.1 to 3.1.1.4) are combined for discussion in this Paragraph. The process of site characterization is linked to the procedures described in Section II of Chapter 2, where sampling distribution design was discussed. In the RI stage, sample design is likely to be influenced by SI data. In turn, these SI results affect the statistical methods at the planner's disposal for collection of site data.

3.2.1. When scoping for the SI, project planners have expectations about the probable location and nature of contamination. By the time a site reaches the RI, some usable information is usually available. In particular, if a contaminant was identified in the SI, planners may have an idea of the mean and standard deviation of contaminant concentrations. These initial estimates assist in devising a statistical sampling design at the RI stage. Two examples of using site data to support sampling design are presented in this Paragraph. These are "hot spot" sampling and geostatistical sampling, the fundamentals of which are presented in Appendices C, J, and R.

3.2.2. A "hot-spot" typically refers to a localized area of high concentration, but is often otherwise poorly defined (e.g., criteria for the size and concentration of hot spots are often arbitrary or not specified). Hot-spots are not uncommon at sites where waste was released in an isolated region, perhaps during a spill. In addition, hot-spots may occur within broader regions with low, but detectable, levels of contamination. One example of this may be when an area was used to process waste disposal over some time and, at times when a shop or operation was cleaning house, a high concentration of waste would be deposited. However, sample concentrations that exceed a regulatory threshold or other decision limit should not be considered to be hot-spots if these concentrations appear to be randomly distributed and will not necessarily be of concern if they represent a small portion of study area and contain a small contaminant mass.

3.2.3. Case study 1 presents an RI application of the hot-spot identification method discussed in Appendix C.

3.2.4. In this instance, professional judgment led to the determination of the size and shape of the hot-spot. The reader is urged to vary *S* and *L* to identify the sensitivity of hot-spot sampling grids to the assumptions.

3.2.5. As stated previously, there is typically some knowledge of contaminant distribution at a site by the time an RI begins. Geostatistics allow an investigator to extrapolate (and interpolate) what is known in one location to other nearby related locations. Its application relies on the fact that, given a known concentration at one location, an adjacent location is likely to have a similar concentration. The greater the distance from the known concentration, the greater uncertainty there is in predicting a concentration at an unsampled location. This situation can be described as a spatial correlation, because correlations are related to how close samples are to one another. Geostatistical methods are described in detail in Appendices J and R.

3.2.6. Case Study 2 illustrates the use of geostatistics for reducing uncertainty in a project. Although geostatistical techniques are more common for RIs than SIs, they may also be used for SIs if sufficient site data are available.

3.2.7. One of the major RI objectives is identifying the distribution of contamination at a site. As useful as geostatistics are in helping with sampling design, they may also be used in interpreting sample data. The geostatistical method known as kriging (Appendix J and R) is an effective method for interpolating site concentration data under conditions where spatial correlation exists. Kriging is a weighted-moving-average interpolation method. The USEPA developed a two-dimensional kriging package, which is useful in providing a fundamental introduction to the technique (Geo-EAS; EPA/600/4-88/033). Kriging as a method of contouring is described in several readily available texts, and typically requires the use of commercially available computer software with kriging options for contouring (e.g., Surfer, EVS).

3.3. Case Study 1—Hot-Spot Identification. The project team attempted to locate a hot-spot resulting from an uncontrolled water release within a larger storage area. The total storage area was approximately 150 by 200 feet. Because the suspected waste was spilled as a liquid, the hot-spot was assumed to be approximately circular. A best estimate of the diameter was approximately 20 feet. The method proceeded in steps as follows:

3.3.1. A circular hot-spot means *S* equals 1.

3.3.2. The radius of the target spot is 10 feet.

3.3.3. The team assigns a value of 0.1 to the acceptable risk of not finding the hot-spot.

3.3.4. Using *S* and β, refer to Table D-1 (or nomographs presented in Gilbert, 1987) to determine that *L*/*G* is 0.55 for a square grid and 0.50 for a triangular grid.

3.3.5. Using the relationship *L*/*G* and the assumed radius of 10 feet, we see that square grid spacing is 18 feet and triangular grid spacing is 20 feet (values are rounded to the nearest foot to reflect the significant figures).

3.3.6. One sample will be placed at each grid node in the storage area, so that a square grid requires 88 samples and a triangular grid requires 75 samples.

3.4. Case Study 2—Using Geostatistics in Project Planning to Reduce Uncertainty and Cost. At a site in the Midwest, project planners were asked to assess a site potentially contaminated with lead at levels exceeding risk-based limits. A SI was conducted using a grid system over areas that were suspected of being contaminated based on historical information.

3.4.1. The project team identified lead concentrations in soil exceeding threshold values in various areas of the site (red circles in Figure 3-1). They were required to move on to an RI/FS to more fully characterize the nature and extent of contamination and develop preliminary estimates of cost for a removal action. Initially, the team intended to collect numerous additional samples on a grid (green circles in Figure 3-1) to more fully delineate the extent of contamination. However, the project geologist suggested the use of geostatistics as a means of reducing the number of samples without increasing uncertainty.

THIS SPACE INTENTIONALLY LEFT BLANK

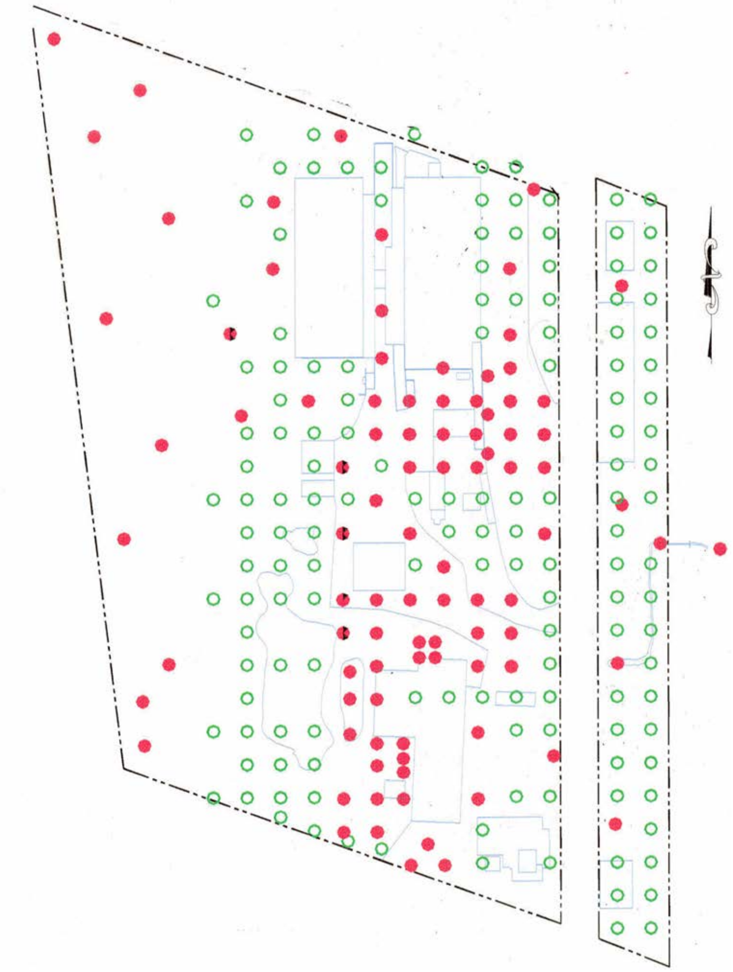


Figure 3-1. Initial Sampling Grid and Proposed New Samples

3.4.2. Geostatistics can predict both the concentration and the uncertainty for an unsampled portion of the study area. In essence, spatial correlations for contaminant concentrations established from the existing data set are used to “extrapolate” sample concentrations and uncertainty for other portions of the study area. Consequently, the team was able to use a geostatistical evaluation to assess the value of collecting additional samples at any given location in the grid. Simply put, the team recognized that in any sampling and analysis system there will be bias and variability, and that estimates of that bias and variability could be made using the existing data. Thus, at any location where the estimate of uncertainty from the geostatistical prediction was less than the uncertainty from sampling and analysis, the team reasoned that there was no value in collecting additional samples.

3.4.3. The final sampling plan required the addition of only seven new sampling points (shown as black circles in Figure 3-2) with associated cost savings of over $12,000.

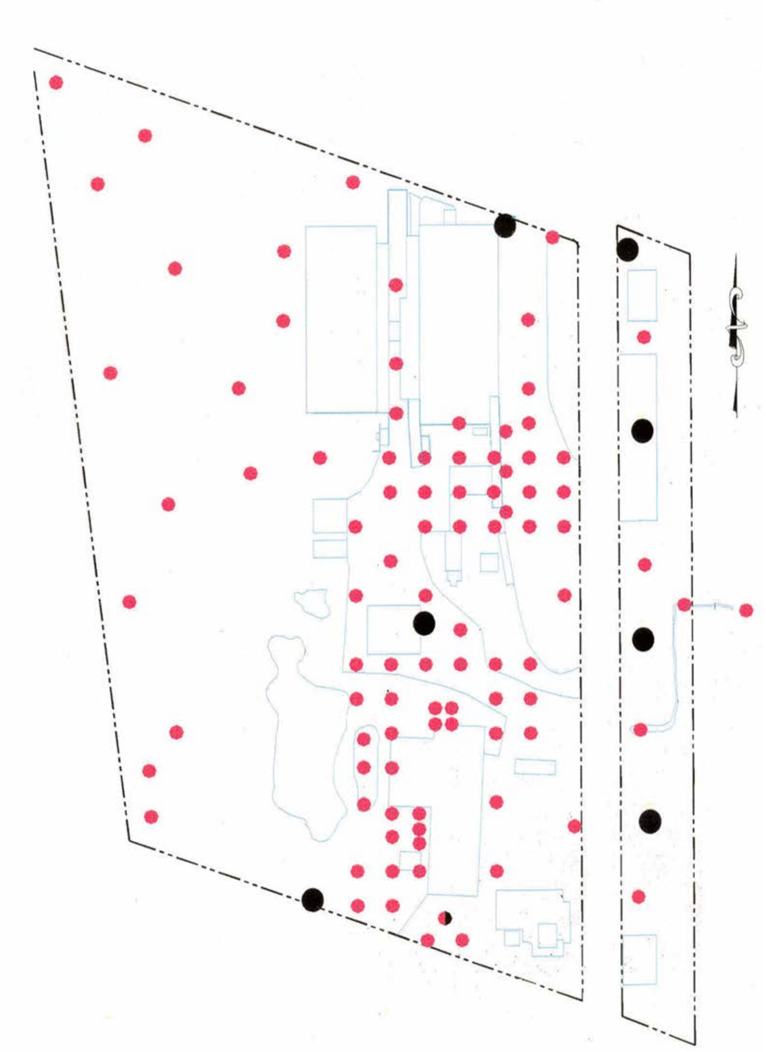


Figure 3-2. Samples Required after Geostatistical Analysis.

Section II
Background Comparisons

3.5. Introduction. Not all chemicals detected at hazardous waste sites originate from site-related activities; for example, metals in soil and groundwater are often present because of natural geological conditions. Similarly, anthropogenic activities unrelated to a site frequently contribute certain organic chemicals (e.g., polycyclic aromatic hydrocarbons [PAHs] or pesticides derived from urban or agricultural sources; EPA SOW No. 788). If site sample concentrations for a specific compound are similar to or lower than background concentrations*, there may be no need to consider potential remedial actions with respect to that

* Background does not mean pristine or unaffected by human activity, especially at sites in heavily industrialized areas.

compound. This determination can be quantitatively defended by use of statistical comparison methods.

3.5.1. The project team should determine the background sampling locations and parameters during the planning stages of the RI. Separating and identifying background sample locations from portions of the study area that have been potentially affected by waste handling activities is an example of stratification. The critical factor distinguishing a background sample from the site lies in understanding where contaminated areas end and natural conditions begin. Such samples may be located upwind, upstream, or upgradient from the waste site. Background data should be drawn from media that physically represent the study area; they should be from the same soil type or geological deposit, same type of surface water system (for example, freshwater versus saltwater; wet season versus dry season), or from the same aquifer as the site data. It is also critical to collect the background samples in substantively the same manner that the site samples are collected (same analytical method, volume of sample, etc). The sampling design and analytical methodology for the background and the site study areas must be similar. For example, erroneous conclusions can result if judgmental sampling is done for the site study area but random sampling is done for the background study area.

3.5.2. Background locations should be in a nearby portion of the region unaffected by site activities. As a caveat, site planners should be skeptical if regulators prefer to limit background sampling to only pristine areas; doing so will potentially result in erroneously concluding that the study area has been adversely impacted by site-related waste handling activities.

3.6. Does Background Soil Differ From Site Soil? The USEPA has developed guidance for addressing whether site soil characteristics differ from background (EPA/540-R-01-003 and EPA/540/S-96/500). The guidance EPA/540-R-01-003 emphasizes the formulation of DQOs in devising background sampling design and subsequent site to background testing. The focus of the cited guidance is only to determine whether site and background soil chemistry differ. It does not establish comparison standards, or levels of background that may replace unnaturally low risk-based clean-up goals.

3.6.1. Fundamentally, the USEPA guidance (EPA/540-R-01-003) identifies two forms of background testing:

3.6.1.1. Background Test Form 1. Tests the null hypothesis that the mean contaminant concentration in samples from the site waste handling area is less than or equal to the mean concentration in background areas.

3.6.1.2. Background Test Form 2. Tests the null hypothesis that the mean contaminant concentration in samples from the site waste handling area exceeds the mean concentration in background areas by more than a specified margin (e.g., by 50 ppm).

3.6.2. Before continuing with this approach, investigators need to be certain that these tests are applied to random sample data sets collected from both the site and background locations. Typically, site sampling may have a component of judgmental sampling, meaning samples were biased to expected contaminated areas of a site. In such cases, the background testing cannot be applied.

3.6.3. The project planning team should establish which form of background testing will be applied at the onset of the RI planning process. In addition, the planning team needs to establish the levels of acceptable levels of error in the decision-making. This will differ from site to site, and will depend on the desires of the project planning team members.

3.6.4. The USEPA guidance also provides examples for the application of test methods that may be applied to the background test forms (EPA/540-R-01-003; Table 3-1). These are:

3.6.4.1. Descriptive Summary Statistics. These (e.g., mean, median, standard deviation, variance, percentiles—see Appendix D) may be used as a preliminary screening tool for comparison with site history and land use activities in the establishment of background. EPA considers these "simple and straightforward [but having low] statistical rigor."

3.6.4.2. Simple Comparisons. These (i.e., greater than maximum) may be used with very small data sets. This approach is not recommended.

3.6.4.3. Parametric Tests. These (e.g., Student *t*-test–see Appendix L) may be used if a larger number of data points is available ($n > 25$). EPA states that parametric tests require approximate normality of the estimated means and recommends that, for smaller data sets, investigators examine data for normality or lognormality in distribution. EPA considers this application statistically robust enough to be used frequently in parametric data analysis.

3.6.4.4. Nonparametric Tests. These (e.g., Wilcoxon Rank Sum Test—see Appendix M) may be used when data are not normally distributed, as rank-ordered tests make no assumption on distribution. Again, EPA considers this approach statistically robust and to be used frequently in background estimation.

3.6.5. The list of methods is not complete, but, by reviewing the appropriate Appendix, users of this Manual may identify the most appropriate statistical method for site application. USEPA guidance leans heavily toward parametric and nonparametric tests, which in turn rely on establishing whether data are normal or lognormal (see Appendix F).

3.6.6. The U.S. Department of the Navy (DON) also developed statistical guidance for evaluating background in soils (UG-2049-ENV). Like the USEPA method, the guidance suggests comparative methods for testing whether site data differ from background. However, DON guidance is unique, in part, because it also relies on geochemical relationships. UG-2049-ENV provides guidance for evaluating the geology of the site and the geochemical characteristics of site soils as they relate to background analyses. The procedures outlined in UG-2049-ENV can be quite useful for USACE projects and are recommended as a resource for additional reading.

3.6.7. This "geochemical method" is often used when reference area data are not available. The method may be used to extract background concentration ranges by evaluating correlated background chemicals using on-site data only (i.e., no background area need be sampled). The key concept is that if the site has not been affected by a release, then only one population exists at a site; if a release has affected the site, then overlapping of different population characteristics would be evident in the data.

3.7. Simple Background Comparison. Investigators are more likely to rely on regional background at the SI stage than the RI. As the text below states, site-specific background is more desirable, but SI project budgets rarely allow for a full background study and such regional comparisons are still useful. Background concentrations are typically not known prior to RI activities, and sampling for background should be scoped in the planning stages of the RI. In some instances, background criteria are available as regulatory limits, as Case Study 3 illustrates. (Although the case study could also apply in an SI [Chapter 2], it is presented here to illustrate the concepts that arise for background comparisons all in one section of this document.)

3.8. Case Study 3—Comparison to Regional Background. Site-specific background concentrations are typically not known prior to RI activities, and sampling for background should be scoped in the planning stages of the RI. In some instances, regional background values may be compared to site data.

3.8.1. Texas has established soil background levels that can be used in the screening process if site-specific background levels are not available. Soil data from one site proposed for redevelopment were compared to Texas background levels. Texas regulation states that if the maximum concentration of the chemical under investigation does not exceed the Texas soil background level, then that chemical is not of concern. The site analytical data were reviewed for quality and applicability. Based on the review, the project team was satisfied that the site analytical data were of sufficient quality for use in evaluating the site. The soil analytical data (in mg/kg) for chromium were:

6.17	4.31	4.38	6.07	5.68
2.86	5.08	4.98	2.22	15.30
4.75	3.56	4.48	3.46	2.63

3.8.2. The maximum concentration for chromium at the site is 15.30 mg/kg. The Texas soil background level for soil is 30 mg/kg. Therefore, chromium would not be a chemical of concern at the site.

3.8.3. As indicated in the USEPA guidance, such a comparison lacks statistical rigor, but is useful for guiding the project planners in the next phase of investigation.

3.8.4. At this stage, the comparison to regional background is merely sufficient to proceed to additional phases of site chromium evaluation.

3.9. Parametric and Nonparametric Tests. In the preceding case study, the regulatory community established background concentrations. It is far more desirable for local background levels to be assessed and applied. Differences related to sample medium, sampling method, or analytical method are less likely to arise in site-specific background data than regional background data. However, the project must be budgeted for a sufficient number of samples to characterize site-specific background conditions; a large number of samples may be required to characterize heterogeneous background media. If the regional background data (e.g., the background data from a very limited site-specific background study) are shown to be statistically different from a waste site, it may also be attributable to differences in water quality or soil types between the site and the location where the regional background data were collected, and not necessarily related to a waste release. Therefore, a thorough evaluation of local background conditions is preferred to the use of regional background levels.

3.9.1. Instructions and guidance for selecting analytical procedures as part of DQOs should be applied to the background data set with the eventual uses of background data in mind. For statistical comparison, background measurements need to be random. In addition, the power of statistical comparison may be greater if the background results are normally or lognormally distributed. Although the distribution of background measurements cannot be guaranteed, either random or systematic sampling of background should be a component of the sampling plan. (Note that given spatial correlation, systematic samples spaced closer than the geostatistical range may not be independent. Sampling methods are addressed in Appendix C.) Once a set of background samples have been collected, comparison methods are applied using the statistical procedures addressed in Appendix M or N.

3.9.2. A random sampling[*] design is typically used to characterize the background study area. Two-sample statistical tests[*] are then typically used to compare the site data set to the background data set. Two-sample tests, described in Appendix M, are summarized in Table 3-1.

[*] Appendix C.

3.9.3. An example of determining COPCs using background population tests is presented in case study 4.

Table 3-1.
Background Population Comparison

Percent Detections in Site Data	Percent Detections in Background Data	Test
0–100	0	No comparison
> 0–100	< 10	Poisson UTL
10–50	10- 50	Test proportions
> 50	> 50	Mann-Whitney test,
85–100	85–100	Student’s *t* test* or Mann-Whitney test

*Student’s *t* test should be used if the distributions of the site and background data sets are normal.

3.10. Case Study 4—Establishing and Comparing Background Concentrations to On-site Data. At a military installation in Utah, samples were collected for metals in soil—seven on site and four at background locations. This case study focuses on chromium. The chromium results were as follows (mg/kg):

SS01	SS02	SS03	SS04	SS05	SS06	SS07	BKG1	BKG2	BKG3	BKG4
4.3	2.7	2.2	3.2	<1	3.6	2.4	1.6	1.8	2.6	1.6

3.10.1. Because the site data had an 85% detection rate, one-half the reporting limit was substituted for each non-detect for the statistical calculations.

3.10.2. Both background and site data were determined to be normally distributed at a 90% confidence level. An *F-test* was used to compare the variance of the background data set to the variance of the site data set. The result of the *F*-test indicated that the variances are equal.

3.10.3. Thus, a two-sample t-test (with equal variances) was used to compare the background and on-site data sets. At the 95% confidence level, the calculated $p = 0.172$. Based on this evidence, a statistical difference between background and on-site data could not be demonstrated at the 95% level of confidence; thus, no further action with respect to chromium was required. Note that, for this simple example, the conclusion of “no further action” is drawn because a statistical difference was not obtained. The power of the test is normally calculated when the null hypothesis is not rejected. Additional investigation would be required if the power was not adequate.

3.11. Upper Tolerance Limits. Upper tolerance limits[*] (UTLs) are sometimes used to determine whether site concentrations are elevated relative to background concentrations. The UTL defines a threshold value for the background data set. (More accurately, it is an upper confidence limit for some percentile of the background data.) Individual site contaminant concentrations are compared to this value. Study area detections that are greater than the background UTL are considered to be indicative of contamination from site-related waste handling activities. Tolerance limits are used in this manner in the USEPA guidance for the statistical treatment of groundwater monitoring data (EPA 530-SW-89-026, EPA 9285.7-09A). However, this approach must be used with caution. In particular, it is often erroneously concluded that site-related contamination exists if a single detection exceeds the UTL. For example, the “95% UTL” is typically used to evaluate site contamination relative to background. If the background and site concentrations are not different from one another, we will be 95% confident that at least 95% of all site measurements will fall below the 95% UTL with coverage of 95%. (For brevity, this is often referred to simply as the “95% UTL.”) Therefore, we would expect a small percentage of site measurements to exceed the UTL, even when overall site contamination is not elevated relative to background. When a large number of samples are taken, we should not definitively conclude that a small number of detections greater than the UTL necessarily indicate site-related contamination.

3.11.1. Furthermore, regulators have criticized the use of UTLs to compare site to background contamination because UTLs do not minimize false negatives but, rather, minimize false positives. In other words, if many detected study area concentrations were greater than the background UTL, this would constitute strong evidence of site-related contamination. This scenario would be unlikely if the site and background concentrations were similar. Alternatively stated, the probability of a false positive—erroneously concluding that the site is contaminated relative to background—would be low. However, if detected site concentrations were less than the UTL, strictly speaking; no conclusion would be possible. This would not be sufficient to demonstrate the absence of site contamination relative to background. If we were to conclude the absence of site-related contamination using the UTL, false negatives could result (i.e., erroneously concluding that site concentrations are not elevated relative to background concentrations).

3.11.2. Because of the problems with tolerance intervals discussed above, two-sample statistical tests are usually preferred (and are typically more appropriate) to compare site and background data sets. It is recommended that UTLs be used only when two-sample tests are not practical (or when the primarily objectives is to demonstrate that site contamination is elevated relative to background contamination). For example, a two-sample statistical test cannot be performed when the site data set is extremely small (when only one or two samples are available for the study area). If a large data set was available for the background study area (e.g., because a “site wide” background study had been done for a prior investigation), then the study area results could be compared to the background UTL.

[*] Appendices G and K.

3.11.3. The UTL background comparison methods are discussed Appendix K. These methods are summarized in Table 3-2

3.11.4. There are parametric UTLs and non-parametric UTLs. The parametric UTL require the data to follow a specified distribution such as a normal or lognormal distribution. (Distribution tests are addressed in Appendices F and J.) As shown in the table above, the proportion of non-detects must be taken into account when selecting an appropriate UTL. (UTLs that rely upon the normality assumption cannot be calculated when a large portion of the data are reported as non-detect.) The nonparametric UTL represents a high-end value in the distribution. The following case study illustrates an example of calculating background UTLs for metals.

Table 3-2.
Calculation of Background UTL$_C$

Percent Detections in Background Data	Type of UTL Calculated
0	No UTL calculated
< 10	Poisson UTL
10–85	Nonparametric UTL
≥ 85 (normal or lognormal distribution)	Parametric UTL

3.12. Case Study 5—Calculating Background UTLs for Metals. At a site in Utah, 56 soil samples were collected across a very large area to determine background concentrations for metals.

3.12.1. Chromium was detected above the detection limit in every sample, so there was no need to substitute for censored values. Manganese was not detected in one sample, and the geochemist elected to substitute one-half the detection limit for the censored value in that sample.

3.12.2. The chromium data were normally distributed and the manganese data were lognormally distributed.[*] Refer to Appendices D, E, and I for a review of these concepts.

3.12.3. For chromium, the 95% UTL was calculated from the sample results using the formula:

$$95\%\ \text{UTL} = \bar{x} + ks\ .$$

3.12.4. For 56 samples, *k* equals 2.032. Chromium results for background had a mean ($\bar{x}$) of 12.7 mg/kg and standard deviation of 5.1 mg/kg, so the UTL was 23.0 mg/kg. For manganese, the log of each sample result was taken prior to the calculation of the UTL. (The

[*] The *Shapiro-Wilk test* (Paragraph F-3) was used to test for normality at the 95% level of confidence.

individual concentrations are not shown.) For the set of log-transformed results, the sample mean and standard deviation were 5.41 and 0.75, respectively. The log UTL for manganese was 6.93 (using the above equation). All comparisons for manganese should occur in "log space" (that is the logarithm of the site manganese maximum would be compared to 6.93). (Alternatively, a minimum variance unbiased estimator of the manganese background concentration could be calculated using the methods described in Appendix E).

3.13. Extended Background Example. This paragraph illustrates the concepts of distributional assumptions presented in Appendix J through a case study.

3.13.1. Suppose surface soil samples (from 0 to 5 feet below ground surface) have been collected at Site A and a background location to evaluate chromium concentrations on site. Table 3-3 presents the analytical results from samples collected at the site and background areas. All chromium concentrations were detected so no proxy concentrations are needed to evaluate the data.

3.13.2. Further, suppose the objectives of this data evaluation are to identify whether chromium surface soil concentrations on site:

3.13.2.1. Exceed regulatory threshold levels.

3.13.2.2. Exceed background concentrations, on the average.

3.13.3. Several statistical tests can be used to make such comparisons. A "one-sample" test can be used to compare the mean site chromium concentration to regulatory risk-based levels (Appendix L). A "two-sample" test can be used to compare the mean concentration of chromium at the site to the mean background concentration of chromium (Appendix M). A background value, such as a UTL, can be estimated for comparisons to individual site concentrations to identify if any one sample has a concentration higher than background. However, before any statistical tests can be done, distributional assumptions must be evaluated for each population (site and background) of data to determine which statistical test is most appropriate. The distributions are evaluated for normality (or log normality) using statistical tests and graphical plots.

3.13.4. Graphical displays are the first approach taken to evaluate the distribution of the data (Appendix J). Histograms, box-and-whiskers plots, and probability plots are all useful in identifying how data are distributed and answering questions such as—are the data symmetrical, what is the range of concentrations, are there any outliers that may unduly influence future distributional tests, do the data seem to follow a normal distribution, and so on. Histograms, box-and-whisker plots, and probability plots for the site and background data are provided in Figures 3-3 and 3-4, respectively.

Table 3-3.
Analytical Results for Chromium at Site A and Background Locations

Site A Sample Location	Top Depth of Sample (ft)	Bottom Depth of Sample (ft)	Chromium Concentration (mg/kg)	Background Sample Location	Top Depth of Sample (ft)	Bottom Depth of Sample (ft)	Chromium Concentration (mg/kg)
SB01	1	2	4.76	BG01	1	2	4.99
SB01	4	5	4.42	BG01	4	5	4.35
SB02	1	2	4.68	BG02	1	2	4.61
SB02	4	5	4.82	BG02	4	5	4.83
SB03	1	2	4.36	BG03	1	2	3.92
SB03	4	5	4.37	BG03	4	5	5.09
SB04	1	2	4.09	BG04	1	2	5.19
SB04	4	5	4.14	BG04	4	5	4.54
SB05	1	2	4.78	BG05	1	2	5.49
SB05	4	5	4.94	BG05	4	5	4.3
SB06	1	2	3.35	BG06	1	2	5.67
SB06	4	5	3.08	BG06	4	5	4.16
SB07	1	2	10.1	BG07	0.5	1	5.41
SB07	4	5	18.5	BG07	2	2.5	4.98
SB08	1	2	10.6	BG08	1	2	5.64
SB08	4	5	4.87	BG08	4	5	4.98
SB09	1	2	10.3				
SB09	4	5	5.51				
SB10	1	2	6.4				
SB10	4	5	4.13				
SB11	1	2	4.96				
SB11	4	5	4.96				
SB12	1	2	4.91				
SB12	4	5	4.89				

3.13.5. These plots have been developed on the basis of the original data and the natural-log transformed data, as it is common that environmental data follow either a normal or lognormal distribution. Other less common transformations, such as the square root or inverse sine transformation, are not applicable in this case study because:

3.13.5.1. Chromium concentrations are continuous (values can be any number within a range of concentrations).

3.13.5.2. Detected chromium concentrations are not rare events to warrant review of the Poisson distribution.

3.13.5.3. Chromium concentrations are not binomially distributed.

3.13.6. Based on just the plots in Figure 3-3, chromium at Site A does not appear to have a normal or lognormal distribution. The histograms for the original data and log-transformed data are not symmetrical, but are skewed. This is confirmed in the box-and-whiskers plots because the mean (the dotted line) is larger than the median (the solid line within the box) and the mean is even larger than the 75th percentile (the top part of the box). (If the data were normal, the mean would be equal to the median.) As the mean is greater than the 75th percentile, this suggests that the mean is influenced by several considerably large concentrations. Outliers (each of point represented by an “x”) predominantly occur only in the upper portion (the top) of the box plots. Lastly, as the normal probability plots for the original data and log-transformed data are not linear, this gives additional evidence that the data are not normal or lognormal.

THIS SPACE INTENTIONALLY LEFT BLANK

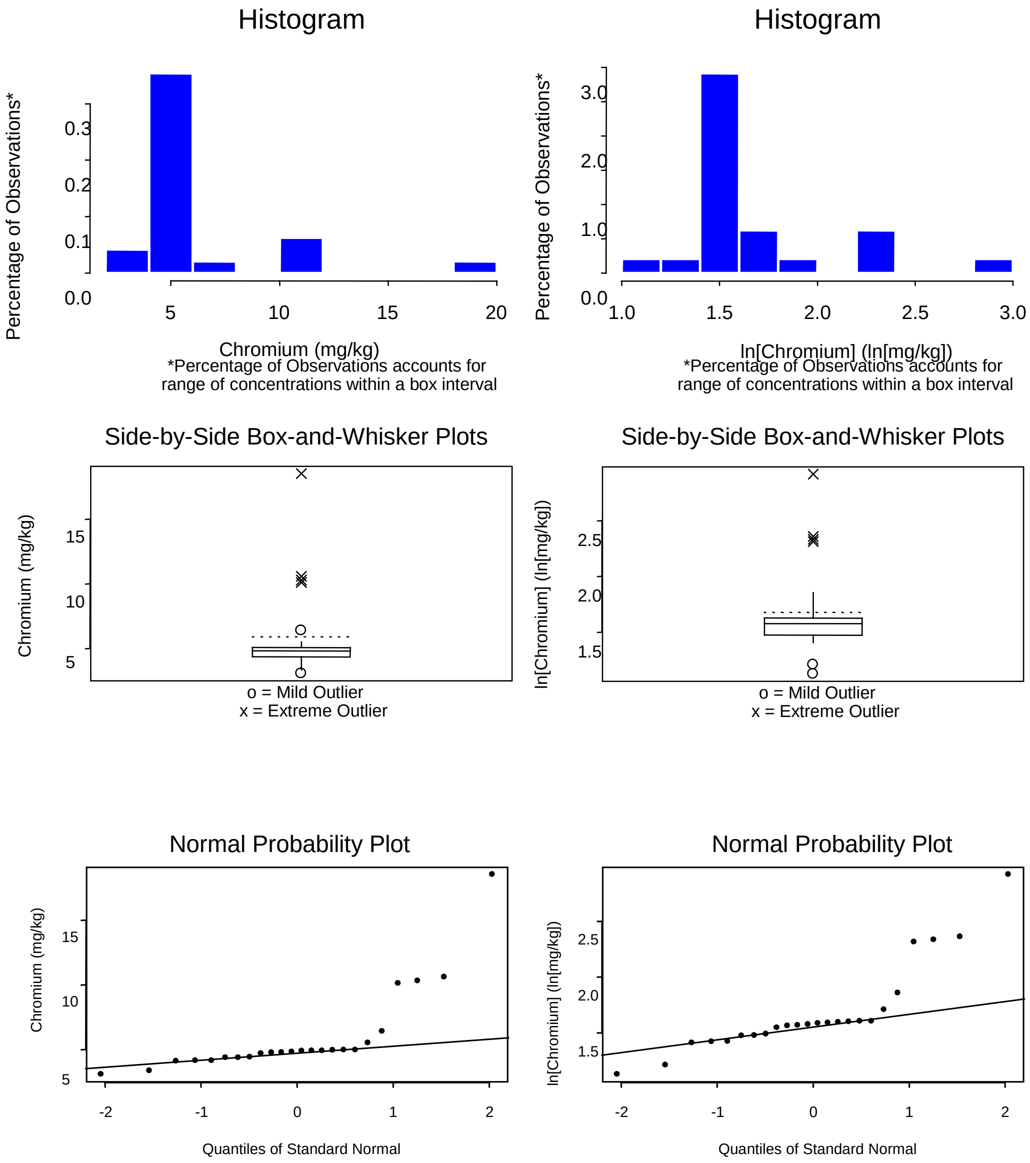


Figure 3-3. Chromium in Site A

3.13.7. The chromium data distributions possess heavier right tails relative to a normal distribution. Note the extreme deviation from linearity (Appendix F) at the right-hand side of each normal probability plot (appearing as a series of points above the straight line). The superimposed line on the normal probability plots illustrates the line that concentrations follow when data are normally or lognormally distributed. This line is related to Filliben's statistic in the sense that it provides a standard to compare the linearity of sample results. For these normal probability plots associated with Site A, it is apparent that the data do not follow a normal or lognormal distribution.

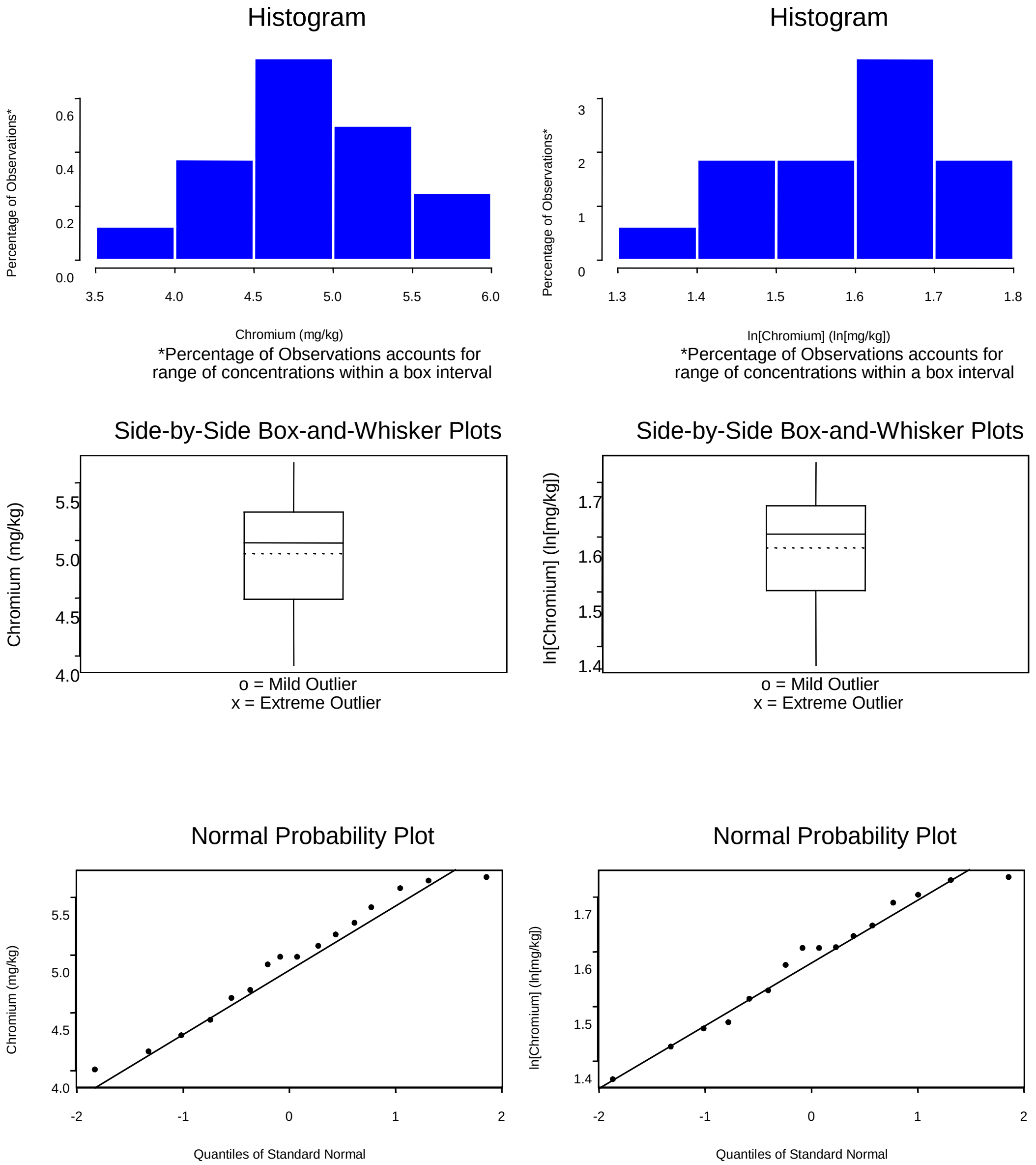


Figure 3-4. Chromium in Background

3.13.8. The plots in Figure 3-4 show evidence that chromium for the background data set appears to follow a normal or a lognormal distribution. The histogram for the original data seems to be symmetrical, though the histogram for the log-transformed data is not as symmetrical. However, histograms can be misleading if the boxes (i.e., concentration intervals) are too large or too small; therefore, another type of plot, preferably a normal probability plot, should be constructed to determine whether the data are normally (or lognormally) distributed.

3.13.9. One of the most powerful statistical methods for testing normality is the Shapiro-Wilk* test. Because the site data set has 24 sample results and the background data set has 16 sample results, this test would be appropriate for evaluating normality and lognormality for both the site and background data sets. The result of the Shapiro-Wilk test is presented in Table 3-4 for chromium at Site A and background based on the original data and log-transformed data. The Shapiro-Wilk test results in either a calculated value of the statistic W or the value p. There is acceptably strong evidence that the data set is not normal when either W or p is small relative to the corresponding acceptance limit for W or p.

3.13.10. For Site A, results of the Shapiro-Wilk test* show evidence that the data do not follow a normal or lognormal distribution (i.e., since the calculated value of W is smaller than $W_{0.01}$, or equivalently, $p < 0.01$, there is less than a 1% chance that the data set is normal, or equivalently stated, there is at least a 99% confidence that the data are not normal). However, for background the results of the Shapiro-Wilk test suggest that the data seem to follow both a normal and lognormal distribution. It should be noted that there is more evidence that background data are normally distributed rather than lognormally distributed, because the value of W and the associated value of p are higher for the original data than for the log-transformed data.

3.13.11. The coefficient of variation* (CV) (*CV*) was estimated for each data set, and is provided in Table 3-4. A *CV* greater than 1 suggests a departure from normality. However, the evaluation of the *CV* is not as reliable as quantitative statistical tests for normality, such as the Shapiro-Wilk test. The coefficient of variation is useful only for identifying obvious departures from normality when *CV* is much greater than 1. Because the sample *CV*s for the site and background data sets based on the original data and the log-transformed data all are less than 1 (as discussed in Appendix F), one cannot conclude the data can be modeled by a normal distribution. Therefore, for these data sets, the *CV* does not provide any useful additional information.

3.13.12. Similarly, to illustrate the relative reliability of various distributional test methods, the Studentized range test* was also performed on the data sets. The results of this test (Table 3-5) indicate that the Site A and background data sets follow normal and lognormal distributions. The range test failed to identify the lack of normality for Site A data. This happened because the data distribution for Site A is asymmetrical and this test does not perform well for asymmetrical distributions. However, according to the test, the background data follow a normal and lognormal distribution. Therefore, the Studentized range test for the background data set is consistent with the Shapiro-Wilk test, the coefficient of variation test, and the graphical plots (e.g., the normal probability and box plots).

3.13.13. Similarly, to illustrate the relative reliability of various distributional test methods, the Studentized range test * was also performed on the data sets. The results of this test (Table 3-5) indicate that the Site A and background data sets follow normal and

* Appendix F

lognormal distributions. The range test failed to identify the lack of normality for Site A data. This occurred because the data distribution for Site A is asymmetrical and this test does not per-form well for asymmetric distributions. However, according to the test, the background data follow a normal and lognormal distribution. Therefore, the Studentized range test for the background data set is consistent with the Shapiro-Wilk test, the coefficient of variation test, and the graphical plots (e.g., the normal probability and box plots).

3.13.14. To summarize, the background data appear to follow both a normal and lognormal distribution, but Site A data do not appear to follow either a normal or lognormal distribution. A dilemma exists regarding the distribution of the background data—is it normal or lognormal? As the log transformation did not appreciably improve the normality of the data set, it would be advisable not to perform the transformation.

3.13.15. If a background value, such as a UTL, and other summary statistics are desired to characterize the background data set, then the assumed distribution should fit the data as much as possible. With respect to this objective, it would be more appropriate to define background as following a normal distribution because the Shapiro-Wilk test shows more evidence of normality than lognormality. Comparing the Shapiro-Wilk test's critical value or associated *p* value from the original data and from the log-transformed data is a reasonable approach for discerning which distribution is more appropriate and has more evidence of following a normal or lognormal distribution.

3.13.16. The first objective for this case study is to determine whether chromium contamination at Site A, on the average, exceeds a regulatory threshold value. As it cannot be assumed that the Site A data set is either normal or lognormal, a nonparametric test (e.g., the Wilcoxon signed rank test for the median as discussed in Appendices H and M) must be used to compare the Site A data to the regulatory threshold.

3.13.17. The second objective is to determine whether chromium exceeds background. Though the background data set could be reasonably assumed to be either normal or lognormal, this assumption could not be made for the Site A data set. As the Site A data set is neither normal nor lognormal, a parametric two-sample test[*] cannot be used to compare the Site A data set to the background data set (for example, to determine if the mean concentration at Site A exceeds the mean background concentration). Both data sets must follow the same distribution to use a parametric test. For example, both the background and site data sets must both be normally or lognormally distributed. As data from Site A does not follow a normal or lognormal distribution, only nonparametric tests such as the Wilcoxon rank-sum test* can be used to compare the Site A and background data sets.

3.13.18. This case study illustrates the value of background data in project decision-making. The application of background data in identifying contaminants for inclusion in the risk assessment is presented in the following section. The data in the preceding discussion may be used as sample data to apply some of the nonparametric tests in Appendix M.

[*] Appendices M and N.

Table 3-4.
Results of the Shapiro-Wilk Test of Normality and Lognormality for Chromium Surface Soil at Site A and Background

Area	Testing for Normality or Lognormality?	Number of Results	*CV*	Shapiro-Wilk Test Statistic, *W*	Critical Values (from Table B-20 of Appendix B)			*p* value for Shapiro-Wilk Test (from statistical software)	Conclusion: Is there evidence that the data are Normally or Lognormally Distributed? Yes/No
					Critical Value based on 0.05 level of significance, $W_{0.05}$	Critical Value based on 0.10 level of significance, $W_{0.10}$	Critical Value based on 0.50 level of significance, $W_{0.50}$		
Site A	Normality	24	0.5687	0.627	0.916	0.930	0.963	<0.0001	No
Site A	Lognormality	24	0.2426	0.791	0.916	0.930	0.963	0.0002	No
Background	Normality	16	0.1093	0.963	0.887	0.906	0.952	0.7177	Yes
Background	Lognormality	16	0.07041	0.958	0.887	0.906	0.952	0.6308	Yes

Table 3-5.
Results of the Studentized Range Test of Normality and Lognormality for Chromium Surface Soil at Site A and Background

Area	Number of Results	Test of Normality (based on original data)			Test of Lognormality (based on log-transformed data)		
		Ratio of Range of Results and Standard Deviation	Critical Values from Table B-21 of Appendix B, assuming a 0.05 level of significance	Conclusion: Is there evidence that the data are Normally Distributed? Yes/No	Ratio of Range of Results and Standard Deviation	Critical Values from Table B-21 of Appendix B, assuming a 0.05 level of significance	Conclusion: Is there evidence that the data are Lognormally Distributed? Yes/No
Site A	24	4.586	(3.308, 4.666)*	Yes	4.400	(3.308, 4.666)*	Yes
Background	16	3.278	(3.01, 4.24)	Yes	3.317	(3.01, 4.24)	Yes

*Critical Values for *n* = 24 are based linear interpolation of critical values from *n* = 20 and *n* = 25.

Section III
Risk Assessment

3.14. Introduction**.** Perhaps more than any other area in the CERCLA project life cycle, assessing site risk relies on statistics. Many of the techniques described in several of the appendices apply in quantifying and assessing risk at a hazardous waste site. The components of a risk assessment discussed in this report are:

a. Identifying contaminants of potential concern (COPCs).

b. Calculating exposure point concentrations (EPCs).

Statistics enter into risk assessment in one additional major area—the calculation of exposure levels. Specifically, a baseline human health risk assessment requires estimation of a reasonable maximum exposure (RME), and a central tendency exposure (CTE). The former relies on 95% upper confidence level (UCL) values for exposure parameters, and the latter on the mean of the exposure parameters. In either case, the exposure parameters are generally provided by EPA guidance, such as the Exposure Factors Handbook (USEPA, 1997). For all practical purposes, the environmental scientist will not need to statistically evaluate these parameters and, consequently, their derivation is not discussed here. However, understanding the concepts presented in Appendix E and K is very useful in deconstructing the data evaluations presented in the Exposure Factors Handbook (USEPA, 1997).

3.14.1. Identification of Contaminants of Potential Concern for Risk Assessment. Not all chemicals detected at a site are typically included in the quantification of risk. Those chemicals retained in the risk assessment are the COPCs. Note that the COPCs are media-specific; COPCs are evaluated for air, surface soil, subsurface soil, groundwater, sediment, surface water, and any other medium sampled in the RI at each site.

3.14.1.1. Chemicals are typically screened against background or other criteria (established by ARARs) and a subset is selected for inclusion in the risk calculations. Some of the screening criteria, other than background levels, include drinking water MCLs, or secondary MCLs, RBCs, and Toxic Substance Control Act (TSCA) values for PCBs (polychlorinated biphenyls) in soil. In addition, inorganics that are essential human nutrients (e.g., iron, potassium, magnesium, sodium, and calcium) may be excluded from the quantitative risk analysis in most cases. (ARARs are identified in the planning stage of the RI.)

3.14.1.2. Both qualitative and quantitative statistical evaluations are frequently performed to identify COPCs. A qualitative evaluation is initially conducted to determine whether select potential analytes of concern can be eliminated from future investigation; a statistical evaluation is subsequently done for a more in-depth look at of contaminants that were not eliminated during the qualitative assessment.

3.14.1.3. For example, for the qualitative evaluation of the data, if a chemical is detected infrequently in the sample data set, and is not considered to be associated with historical waste handling at a site, it may be screened out as a COPC. However, it is essential to use site-specific information before discarding such a chemical, as infrequently detected compounds may also represent hot-spots, depending on the sampling strategy used at the site. For every chemical detected at least once, the maximum detected concentration is compared to the chemical- and medium-specific screening criterion. Chemicals with higher concentrations than their criteria are generally retained for quantitative evaluation in the risk assessment.

3.14.1.4. Contaminants that lack ARARs (usually because toxicity information does not exist) are retained as COPCs in the risk assessment and discussed in the uncertainty section of the report. One-sample tests for contaminants where the maximum exceeds the risk-based screening limit may be used to determine whether the mean is statistically less than the screening limit, even though a single value exceeds the screening limit. Anthropogenically derived contaminants (such as PAHs) that occur at concentrations below background levels are still retained in the risk assessment if they exceed ARARs. If the risk assessment indicates that such contaminants are a primary contributor to total risk at a site, then a quantitative statistical comparison with background (e.g., using appropriate two-sample statistical tests) would be done and the results would subsequently be discussed in the risk characterization at the end of the assessment.

3.14.2. Calculating Exposure Point Concentrations. For risk assessment, means and standard deviations are typically calculated as the basis for EPCs and as the basis for deriving UTLs for the background comparisons. However, the mean and standard deviation will frequently be inappropriate measures of central tendency and dispersion when the data are not normally distributed or a large portion of the data consists of non-detects. Under these circumstances, means and standard deviations should not be used to perform statistical evaluations. Before statistically valid means and standard deviations can be calculated, tests for normality should be conducted and non-detects must be appropriately addressed.

3.14.2.1. The EPC is used to calculate a COPC’s carcinogenic risk and non-carcinogenic hazard index. It represents the concentration a receptor is likely to encounter. The USEPA requires the EPC to be a conservative estimator of central tendency—the 95% upper confidence limit (UCL) of the sample arithmetic mean concentration (OSWER 92-856-03, EPA 68-W0-0025). The 95% UCL is the concentration that, when calculated repeatedly for randomly drawn samples, equals or exceeds the true mean 95% of the time.

3.14.2.2. Calculating rigorous, statistically valid 95% UCLs requires that data be distribution tested and that non-detects be treated properly. Procedures for this are provided in Appendix H. Some of the older (pre-2000) RCRA and CERCLA guidance for calculating the UCL are outdated (and hence, are not recommended); modifications and updates are provided with the goal of improving scientific defensibility. Appendix K presents methods to calculate UCL_S.

3.14.2.3. Calculating EPCs at a CERCLA site brings together many of the statistical procedures described in the attached Appendices. The correct steps are, in general, as follows

3.14.2.3.1. Identify the nature of the censoring limit and the proportion of censored values and substitute proxy values as directed in Appendix H.

3.14.2.3.2. Identify outliers as discussed in Appendices I and J.

3.14.2.3.3. Perform distribution testing as detailed in Appendix F.

3.14.2.3.4. Depending on the outcome of these steps, calculate the UCL as directed in Appendix K.

3.14.2.4. Unfortunately, there are many pitfalls along the way, and this process does not always lead to a simple result. In part, this is attributable to the use of or adherence to older USEPA guidance. In particular, USEPA guidance for substituting for censored data is addressed in many separate risk assessment documents. In earlier documents, substituting one-half the detection limit is supported. Appendix H provides insight on the deficiency in this approach. In addition, even if the risk assessor has performed all of the statistical procedures, USEPA guidance for EPCs states that if a 95% UCL exceeds the maximum value of a compound detected at a site, the maximum should be substituted. This has the dissatisfying attribute of being completely ad hoc, giving rise to unquantifiable and unacceptable uncertainties for risk assessment decisions.

3.14.3. Uncertainty Quantification. A required element in a baseline human health risk assessment is to evaluate uncertainty for decisions. Statistical techniques alone will be unable to account for all sources of uncertainty in a risk assessment and a qualitative approach is normally taken. For example, there will be uncertainty in the risk assessment for analytes for which toxicity data do not exist, and the quantification of such uncertainty is not possible.

3.14.3.1. In risk assessment, uncertainty stems primarily from the following three sources.

3.14.3.1.1. Errors in the estimate of contaminant concentration.

3.14.3.1.2. Errors in the estimate of toxicity.

3.14.3.1.3. Errors introduced by large numbers of assumed values in the risk assessment formulations, which are by definition and intent very conservative.

3.14.3.2. In practical terms, there is little that can be done about the uncertainty in estimates of toxicity. The studies upon which toxicity data are based are taken “as is” simply because of the scarcity of available studies. Uncertainty in the assumptions employed in the risk assessment can sometimes be addressed, but only to a limited extent. An example for how the uncertainties listed in subparagraph 3.14.3.1.3 were taken into account is presented in Case Study 6.

3.14.3.3. Most statistical evaluations implicitly assume the absence of bias. The uncertainty predominantly depends on the distribution of field measurements. Even in the case of risk screening, as demonstrated in Chapter 2, we have seen that it is possible to qualitatively assess the uncertainty of individual sample/analytical results before comparing those results to fixed threshold values using analytical QC information. For example, QC data can potentially be used to identify the direction of bias and to estimate the magnitude of the bias associated with a set of analytical results. This is illustrated in Case Study 6. It is also possible to make similar estimates of variability which may affect decision-making, as illustrated in Case Study 7.

3.14.3.4. The error introduced into the risk assessment by the uncertainty associated with each of the various assumptions and reference values is more likely multiplicative rather than additive, such that the calculated risk is conservative to an extraordinary degree. Consider, for instance, some components of a soil dermal absorption scenario. The risk assessor calculates an EPC, which represents the 95% UCL of the mean. Then, the skin area exposed to the contaminant is based on an upper 95% confidence level of all the U.S. adult population from EPA OSWER 92-856-03. These are combined with, say, the default average exposure duration and frequency values which, again, are upper estimates from some population. Combining all of these upper estimates results in a risk evaluation that has a far higher confidence than 95%. The Risk Assessor and Project Manager are encouraged to identify every opportunity to use site-specific values in place of assumptions in risk assessment to reduce uncertainty in the results and, thus, more appropriately apply the limited remediation resources available.

3.14.3.5. One method for estimating the true mean and distribution of risk estimates is to use the recommended RME and CTE values of exposure parameters. This methodology is recommended in Risk Assessment Guidance for Superfund (RAGS). The result of looking at each input parameter using the CTE is to provide an estimate of risk near the mean of the estimated exposure scenario. The RME is considered to represent an upper estimate of site risk. An alternative method of quantifying the range in risk estimates is to use Monte Carlo simulations.

3.15. Case Study 6—Refining Risk Assessment Assumptions.

3.15.1. A risk assessment was to be done as part of a RCRA Facility Investigation (RFI) at a steel mill in Pennsylvania. The project team approached the EPA Remedial Project Manager (RPM) regarding using site-specific assumptions for some of the exposure factors in the risk assessment calculations. This was possible because the facility maintained excellent records of employee longevity, promotion, and work assignments. For this case study, the focus is on site-specific estimates of exposure duration, which enters into quantification of risk.

3.15.2. Under the assumptions given by the EPA for the worker exposure scenario in OSWER 92-856-03, the risk assessor is to assume that a given worker will be exposed for a period of 25 years. However, by reference to detailed employee records for the facility, the project team was able to demonstrate concretely on a facility-specific, job-specific, and location-specific basis, the actual average lifetime exposure duration for the various site areas under study. Employing these actual values, which were approximately 3 to 5 years rather than 25 years, greatly reduced the exposure duration. More importantly, the site-specific value reduced the uncertainty in the calculated lifetime risk. Using this lower value allowed the steel mill owner to limit the number of site areas proceeding to the Corrective Measures Study phase of the project.

3.16. Case Study 7—Direction and Magnitude of Bias. As part of a property transfer in Baltimore, Maryland, the project team was asked to estimate reserves that the seller would have to put in escrow against the potential need for site clean-up, before the seller would accept transfer of the property. For this case study, petroleum hydrocarbon contamination will be discussed.

3.16.1. The project team decided to divide the relatively small site into four quadrants and collect one composite sample from each to assess the potential need for remediation in each quadrant. The analytical results obtained from the laboratory were as follows:

Quadrant	Result (mg/kg)
1	1200
2	101
3	756
4	138

3.16.2. With the state's action level set at 100 mg/kg, it appeared that the seller would be required to reserve funds against a potential soil removal for the entire site. However, a review of the quality control data associated with the analytical results displayed significant potential bias.

3.16.3. A normal calibration curve was developed for the gas chromatograph used in the analysis that met method criteria for linearity. The laboratory then analyzed an Initial

Calibration Verification (ICV) using a standard from an alternative source from that employed in the calibration. The ICV was essentially a blank spike set at the midpoint of the calibration curve. The result of this analysis was a percent recovery (%R) of 168%, which was within the acceptance limits provided with the standard by the manufacturer.

3.16.4. However, in its simplest form this QC result indicates that if the laboratory introduced the equivalent of 100 mg/kg of total petroleum hydrocarbons (TPH) into the analytical system, they would get a reported result of 168 mg/kg. This observation, applied to the results reported for the site, removed two of the four quadrants from further consideration, reducing the required reserves by half.

Section IV
Probabilistic Risk Assessments Monte Carlo Simulations

3.17. Introduction**.** The implementation of probabilistic risk assessment for environmental projects is beyond the scope of this document; however, a brief overview of the procedures is presented here. Monte Carlo simulation, the most common technique used for probabilistic assessments, is a statistical technique in which outcomes are produced using randomly selected values for input variables that possess a range of possible values. In some cases, a known probability distribution can be assigned to each input variable. By repeating the calculation many, many times, Monte Carlo simulations create a population of results representing (in theory) the full range of possible outcomes and the likelihood of each. For example, when Monte Carlo simulation is used in risk assessment, risk is expressed as a distribution of possible values rather than a single point value.

3.17.1. There are two major practical limitations to the application of Monte Carlo simulations in general: i) it can be costly, and ii) few people are sufficiently qualified to do it. The EPA has also written a guidance document for probabilistic risk assessment titled RAGS Volume 3 Part A: Process for Conducting Probabilistic Risk Assessment (EPA 540-R-02-002) available at http://www.epa.gov/oswer/riskassessment/rags3a/index.htm. An EPA Region 3 publication (EPA 903-F-94-001) identified several technical limitations that preclude the Agency from relying on Monte Carlo simulations: (http://www.epa.gov/reg3hwmd/risk/human/info/guide1.htm).

3.17.1.1. Software is unable to distinguish between measurement variability and lack of knowledge. Some input parameters are for well-described differences among individuals—these differences are variability. Other factors, such as frequency and duration of trespassing, are simply unknown, and assuming a distribution for them is ad hoc. But the simulated distribution of unknowns is presented in computer output as variability. The accuracy of the distributional assumptions limits the accuracy of the simulation.

3.17.1.2. Software is unable to account for sample dependency (e.g., spatial and temporal correlations for sample locations). However, this limitation also applies to all classical statistical methods (e.g., the methods predominantly discussed in this document and in EPA environmental statistical documents such as the QA-G4 and GA-G9 guidance documents). In classical statistics, the assumption of independence highly influences the applicability of a technique—the same limitation applies here.

3.17.2. In most statistical evaluations (excluding geostatistics), environmental scientists are resigned to the limitations of classical statistics for environmental data. The same is true for Monte Carlo simulations. Though Monte Carlo simulations require sample independence, the approach can be advantageous. The primary advantage is that it accounts for a range of input values and outputs a range of outcomes (such as risk values) with associated probabilities. Although a Monte Carlo approach is currently not recommended or required by the EPA, the approach may be beneficial for some projects. There are applications of such simulations. Moreover, future scientists may learn how to overcome some of the limitations and eventually develop reasonable and inexpensive computer applications.

3.17.3. Applications of Monte Carlo simulation are more prevalent in groundwater modeling than any other current environmental application. Case Study 8 shows how a Monte Carlo simulation of groundwater contamination was used to perfect a remedy.

3.18. Case Study 8—Monte Carlo Simulation in Remedial Alternative Selection.

3.18.1. Monte Carlo analysis was coupled with decision tree analysis for a study site in Nebraska where the groundwater was contaminated with trinitrotoluene (TNT). The extent of TNT contamination was characterized during an RI. Three pump-and-treat alternative remedial actions were developed for the FS. The maximum concentration of TNT remaining in the saturated zone at the end of each alternative project lifetime was determined stochastically using a Monte Carlo model. The Monte Carlo model randomly generated values for site information for initial mass concentration, hydraulic conductivity, and retardation coefficient. Then these randomly generated fields were sampled and the output was combined into sets or ensembles. Probability functions were fitted to the output ensembles with the maximum simulated TNT concentrations. Because each of the treatment alternatives was associated with a different set of possible maximum concentrations, the Monte Carlo simulation made it possible to identify the optimal alternative quantitatively by analyzing the output ensembles for each alternative.

3.18.2. Applying Monte Carlo simulations requires the technical support of a specialist in this area; detailed methodologies are beyond the scope of this Manual. The technique does rely on the power of randomly generated data sets and the optimization of conditions based on the simulation.

CHAPTER 4

Remedial Design and Remedial Action

4.1. Introduction. During the RD/RA phase, engineers develop detailed designs for remedial actions, construct remediation systems, and operate and monitor sites with long-term remedies in place. The term remedial system is defined here in a broad sense; it includes removal actions and capping as well as more active treatment systems.

4.1.1. A number of statistical approaches that are applicable for prior stages of a project's life cycle are also applicable for the RD/RA. This Chapter will address environmental statistical applications for the RD/RA that have not been highlighted for the PA, SI, or RI/FS. In this Chapter, we consider adaptive sampling plans for removal actions and groundwater monitoring and trend analysis.

4.1.2. Although groundwater is most commonly subject to long-term monitoring, the same tools can be used to monitor and optimize remedial systems for other environmental media or demonstrate achievement of site closure criteria.

4.2. Comparisons to ACLs and MCLs. Confirmation sampling is often performed for the RD/RA and would typically entail one-sample statistical tests. These would be the same types of tests that would be conducted during the SI and RI, only the nature of the decision limits would differ (e.g., the decision limits for the RD/RA would be "cleanup goals" rather than the risk-based screening concentrations as in the SI).

4.2.1. As an example, consider data collected at a landfill. If a statistically significant difference is observed between upgradient and downgradient concentrations, a compliance monitoring program must be put into place. According to RCRA regulations, analysis of Appendix IX list constituents is required. Assuming that a release is confirmed, the facility must demonstrate that the release does not present a health or environmental risk. Generally, this entails comparing analytical results to fixed threshold values, called Alternate Concentration Limits (ACLs), which are often established in a jurisdiction-specific fashion. An alternative approach is to compare site data to MCLs. In the first case, tolerance or confidence intervals are recommended. In the second case, the tolerance limit is the preferred method.

4.2.2. An appropriate one-sample statistical test is to determine whether contamination exceeds the decision limit (e.g., an MCL). For example, if a set of measured contaminant concentrations is normal, a one-sample *t*-test could be used to compare the mean concentration to the decision limit. However, a reliable comparison using a one-sample test will not be possible if the data set is small (e.g., consists of only three points). If normality of the data set can be assumed, a conservative approach would consist of calculating an UTL and comparing it to the decision limit. If the UTL were less than the decision limit, there would be strong evidence that site contamination does not exceed the decision limit. However, do not conclude that there is a contamination problem when the UTL exceeds the decision limit. To avoid false positives, when the UTL exceeds the decision limit, additional data should be collected to do an appropriate one-sample statistical test.

4.2.3. The confidence limit approach is used for comparisons to ACLs based on background data, whereas the tolerance limit approach is used when the comparison criteria are health-based and the comparisons are in relation to MCLs or health-based ACLs. The tolerance limit approach is more conservative than the confidence limit approach in that the UTL must be less than the MCL. However, Gibbons (1994) has pointed out the following.

4.2.4. Because at most four independent samples will be available during semiannual monitoring, the 95% confidence, 95% coverage tolerance limit is approximately five standard deviation units above the mean concentration. In light of this, even if all four semiannual measurements for a given compliance are well below the MCL, the tolerance limit will invariably exceed the MCL or health-based ACL and never-ending corrective action will be required.

4.2.5. Thus, special care must be taken in the design of compliance monitoring programs to ensure that the facility is not caught in the kind of regulatory trap described above.

4.2.6. In addition to one-sample statistical tests, multi-sample statistical tests can be appropriate for the RD/RA to perform comparisons with background values. Since long-term monitoring is commonly performed for groundwater during the RD/RA, Figures 4-1 through 4-5 summarize the types of one-sample and two-sample statistical tests that would be used for groundwater monitoring.

Section I
Groundwater Monitoring and Optimization Trend Analysis

4.3. Introduction. Monitoring remedial systems have significant, long-term costs. It is not difficult to anticipate that, over the course of 10 to 20 years, substantial economic resources available for environmental programs at military installations will be in long-term monitoring of sites actively under remediation or sites that require long-term monitoring. Project planners should ensure that these monitoring systems are optimized, and that they provide the necessary information at the least possible cost. Likewise, where active remediation is ongoing, optimization is important to minimize economic impacts to the facility. While optimization is desirable, compliance is mandatory, and at most installations, groundwater monitoring is required under various permits or consent agreements. This section reviews various methods of assessing groundwater systems over time with a view to both detection and compliance, and optimization.

4.4. Detection and Compliance Monitoring. Detection monitoring is a means of identifying whether a regulated hazardous waste site is releasing hazardous materials into the environment. Compliance monitoring entails the repetitive, periodic sampling and analysis of a select set of monitoring locations for compliance with a fixed set of standards or requirements. The standards to which analytical results are compared are generally specified in regulations, permits, or consent agreements.

4.4.1. In detection monitoring, the results of sampling and analysis from a location that has recorded a release are compared to measurements from an unaffected or background location. In the case of groundwater monitoring, this generally entails selecting one or more monitoring wells upgradient of the site and selecting a representative set of downgradient monitoring wells. If the difference between the two sets of results is statistically significant, the owner is usually required to begin compliance monitoring to investigate how the release is occurring and to remedy the situation. These statistics fall into the category of hypothesis tests, specifically two- or multiple-population tests, and are addressed in Appendices M and N.

4.4.2. The selection of the statistical approach is generally open to discussion with regulators and the final determination will depend upon many factors. In general terms, the simplest approach (consistent with the requirements of local jurisdictions) is the best approach. For example, for detection monitoring, a two-sample *t*-test could potentially be used to compare upgradient (background) to downgradient (site) contaminant concentrations. Under the best of circumstances, a straightforward, parametric *t*-test would suffice; however, in practical terms, it is rare that environmental data meet all of the conditions that would make such a straightforward approach viable. And, in fact, by the time Figure 4-2 was published in EPA 530-SW-89-026, the use of the *t*-test had been largely discredited for this application because it failed to adequately control false positives when multiple site and background comparisons are required. Clearly, as of the time of its publication, the 1989 guidance recommended the use of ANOVA techniques (essentially a generalization of the two-sample *t*-test), and, to a lesser extent, alternatives such as tolerance intervals, prediction intervals, and control charting. By 1992, with the publication of Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities—Addendum to Interim Final Guidance (EPA 68-W0-0025), a somewhat different statistical approach was highlighted. Preferences had shifted further with the use of intervals and resampling strategies receiving much greater attention. By 1994, when Gibbons published Statistical Methods for Groundwater Monitoring, ANOVA techniques had largely fallen out of use, replaced by prediction intervals with resampling strategies that have become, in some cases, very complex. This statistical approach currently represents what might be called the state-of-the-art for groundwater.

4.4.3. The alternative approach of using control charts has not gone altogether out of favor, however. A control chart is a type of plot (using data from a particular monitoring well) of some function of concentration (e.g., the mean concentration) versus time. The various statistical tests previously discussed are based on one of two possible approaches for detection monitoring. With the exception of the control chart approach, each new downgradient result is compared to the history (or historical data set) of upgradient results. These types of comparisons are called interwell (literally, “between well”) comparisons. A potential flaw in this approach is that it assumes the only variable that can make a difference between the upgradient and downgradient results is the intervening waste management unit. In reality, there are a number of other possible influences and, for this reason, intrawell (literally, “within well”) comparisons are still considered quite useful in groundwater monitoring applications. The classic method of performing these intrawell comparisons is with control charting. The two types of control charts normally employed for these purposes are the Shewart and cumulative summation (CUSUM) control charts, which are often combined in normal use.

4.4.4. Figures 4-1 through 4-5 present flow charts showing the options available and guidance on option selection. However, the decision regarding the type of statistical analysis program to employ should be made as part of the DQO development process for the monitoring effort. It is strongly recommended that the Project Manager involve a statistician in this process.

4.4.5. Case study 1 provides an example in which multiple techniques are used to assess groundwater monitoring data. Case study 2 provides an example of using a combined Shewart/CUSUM method to identify a release at a site.

4.5. Case Study 1—Groundwater Monitoring. At a manufacturing facility in Virginia, a long-standing tetrachloroethene (PCE) plume is being hydrologically contained and treated with a combination of vapor extraction and groundwater pump-and-treat. The facility has been engaged in long-term monitoring for over 20 years and uses a variety of techniques to assess permit compliance. Sample statistics allow the facility to determine whether remediation at the site is causing reductions in PCE concentrations. Table 4-1 presents an example of summary statistics and testing results in a fashion that is easily understood for both compliance and detection monitoring.

4.5.1. For compliance monitoring at wells with known past contamination (MW1 to MW4), increasing or decreasing statistical trends were determined at the 90 and 95% level of confidence, respectively, as negotiated with state regulators at the site.

4.5.2. Trend analyses, control charts, and tolerance limits are being used for the four wells under the category “Comp” and for the three wells under the category “Trend.” Typically, differing DQOs would be set for compliance and detection wells and only one set of statistical tests would be performed. However, the regulatory negotiations at this site mandated identical tests for both types of wells. (This example demonstrates an opportunity for improving past negotiated monitoring with regulators.)

4.5.3. Additionally, the number of detections greater than the "tolerance limit" is specified for each well. The 95% UTL is constructed from a set of background wells, also as determined in the site permit at time of negotiation with regulators. Because there is background contamination the following case study provides an example of using a combined Shewart/CUSUM method to identify a release at a site.

Table 4-1.
Groundwater Monitoring Data for Case Study 1

Identification				**Descriptive Statistics**				**Trend Significance**		**Excursions?**	
Class	**Well**	***n***	**Avg**	**Med**	***s***	***W***	**MK**			**Control Chart**	**Tolerance Limit**
								95%	**90%**		
Comp.	MW1	46	5595.0	5610.0	982.0	Yes	No	Up	Up	None	3
	MW2	44	62.3	67.2	21.5	Yes	No	Down	Down	None	None
	MW3	40	1295.0	1198.0	367.8	No	No	Down	Down	None	None
	MW4	47	133.8	133.7	22.3	Yes	No	Down	Down	None	None
Detect.	MW5	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW6	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW7	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW8	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW9	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW10	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW11	16	0.369	0.4	0.307	Yes	No	None	None	None	None
	MW12	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW13	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW14	16	0.0	0.0	0.0	N/A	N/A	None	None	None	None
	MW15	16	0.039	0.0	0.088	No	No	None	None	None	None

Notes:	Comp	Compliance
	n	Number of samples
	Avg	Sample mean
	Med	Sample median
	s	Sample standard deviation
	W	Normal according to Shapiro-Wilk test at 95% confidence?
	MK	Seasonality according to Mann-Kendall test at 95% confidence?

THIS SPACE INTENTIONALLY LEFT BLANK

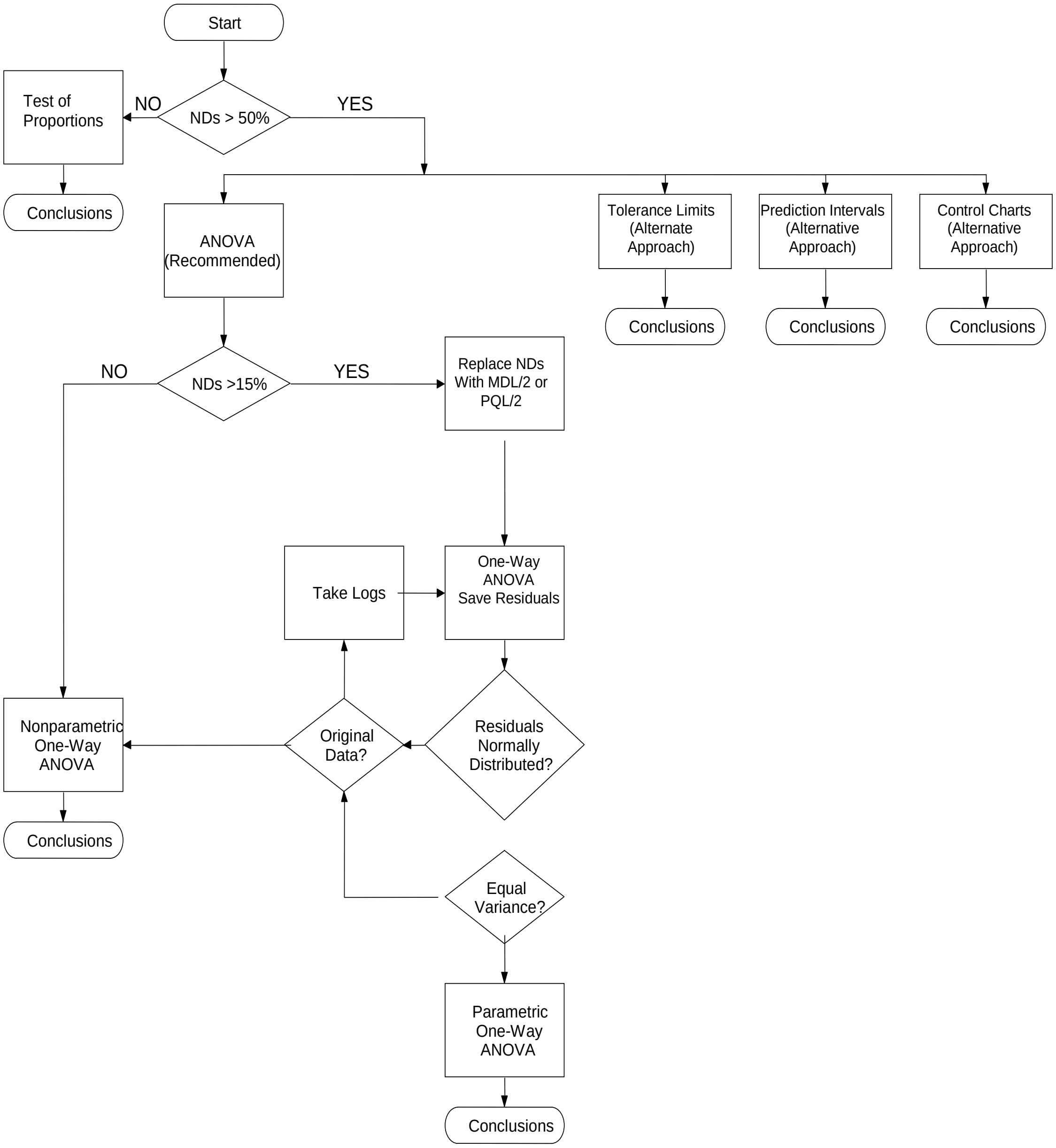

Start
NDs > 50%
NO
YES
Test of Proportions
Conclusions
ANOVA (Recommended)
Tolerance Limits (Alternate Approach)
Prediction Intervals (Alternative Approach)
Control Charts (Alternative Approach)
Conclusions
Conclusions
Conclusions
NO
NDs >15%
YES
Replace NDs With MDL/2 or PQL/2
Take Logs
One-Way ANOVA Save Residuals
Nonparametric One-Way ANOVA
Original Data?
Residuals Normally Distributed?
Conclusions
Equal Variance?
Parametric One-Way ANOVA
Conclusions

Figure 4-1. 1989 EPA Decision Tree for Groundwater Monitoring

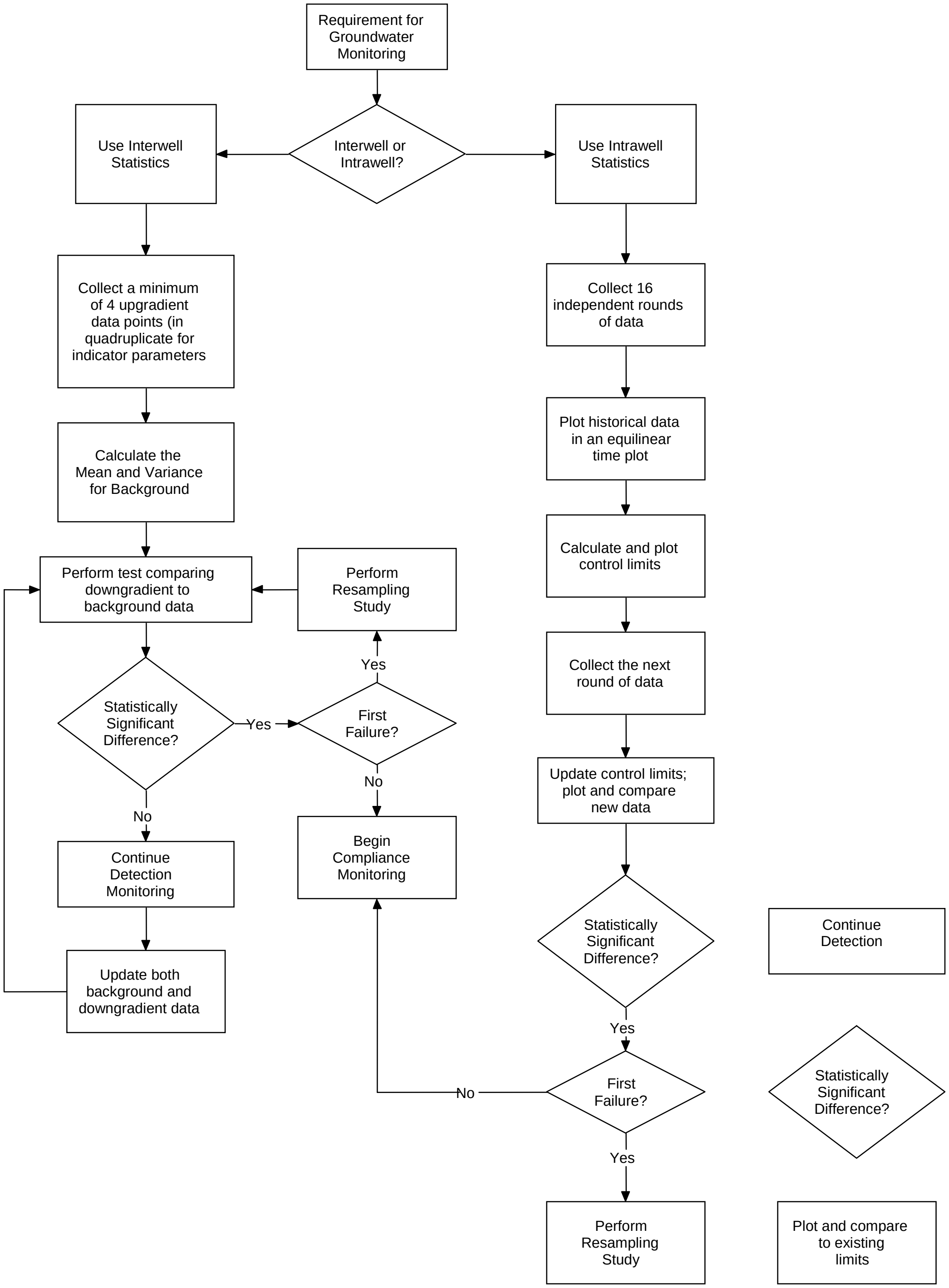

Requirement for Groundwater Monitoring
Interwell or Intrawell?
Use Interwell Statistics
Use Intrawell Statistics
Collect a minimum of 4 upgradient data points (in quadruplicate for indicator parameters
Calculate the Mean and Variance for Background
Perform test comparing downgradient to background data
Perform Resampling Study
Statistically Significant Difference?
Yes
First Failure?
Yes
No
No
Continue Detection Monitoring
Begin Compliance Monitoring
Update both background and downgradient data
Collect 16 independent rounds of data
Plot historical data in an equilinear time plot
Calculate and plot control limits
Collect the next round of data
Update control limits; plot and compare new data
Statistically Significant Difference?
Yes
First Failure?
No
Yes
Perform Resampling Study
Continue Detection
Statistically Significant Difference?
Plot and compare to existing limits

Figure 4-2. Statistical Decision Tree with Options for Groundwater Monitoring-Part 1.

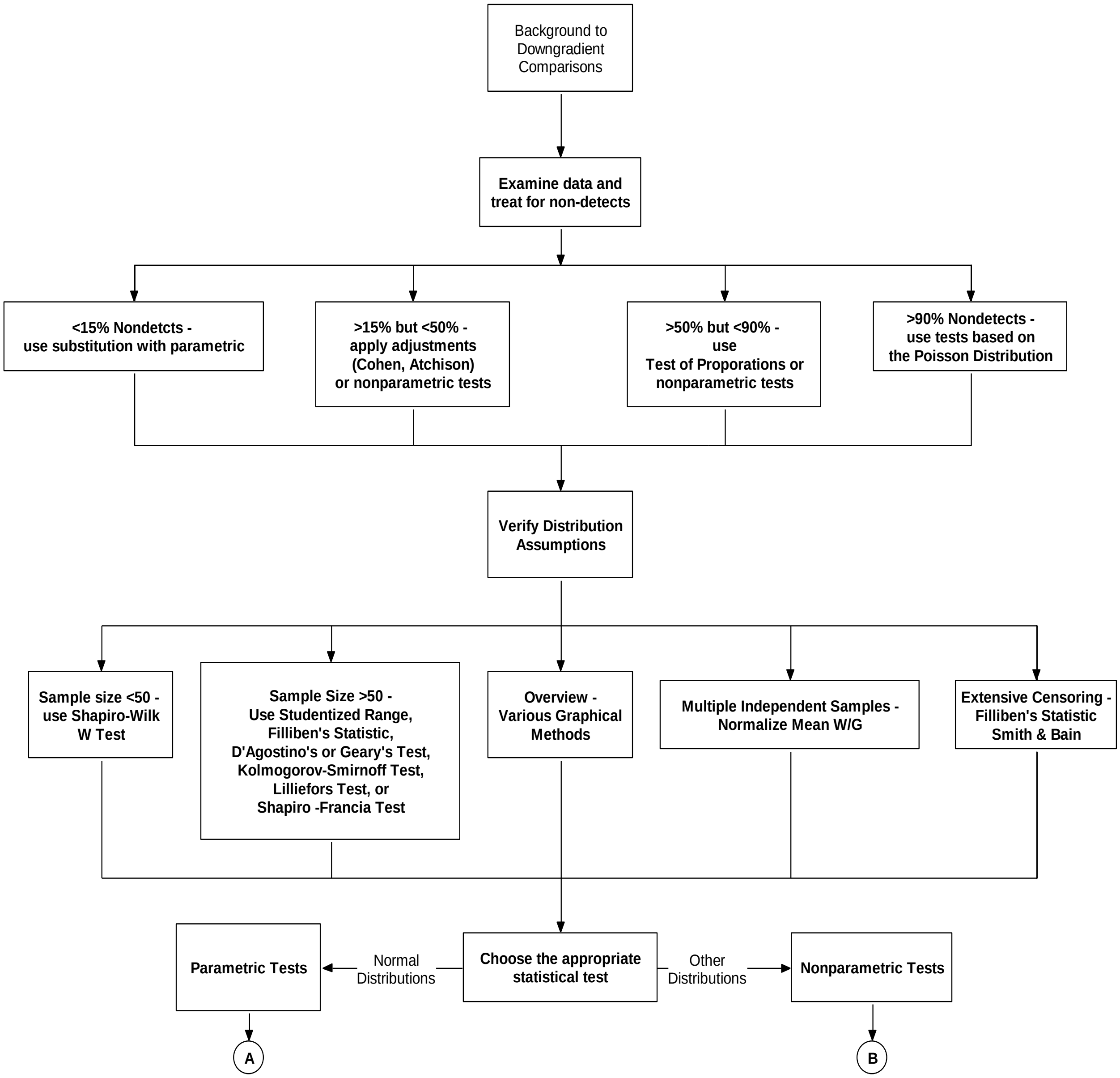

Background to Downgradient Comparisons
Examine data and treat for non-detects
<15% Nondetcts - use substitution with parametric
>15% but <50% - apply adjustments (Cohen, Atchison) or nonparametric tests
>50% but <90% - use Test of Proporations or nonparametric tests
>90% Nondetects - use tests based on the Poisson Distribution
Verify Distribution Assumptions
Sample size <50 - use Shapiro-Wilk W Test
Sample Size >50 - Use Studentized Range, Filliben's Statistic, D'Agostino's or Geary's Test, Kolmogorov-Smirnoff Test, Lilliefors Test, or Shapiro -Francia Test
Overview - Various Graphical Methods
Multiple Independent Samples - Normalize Mean W/G
Extensive Censoring - Filliben's Statistic Smith & Bain
Parametric Tests
Normal Distributions
Choose the appropriate statistical test
Other Distributions
Nonparametric Tests
A
B

Figure 4-3. Statistical Decision Tree with Options for Groundwater Monitoring-Part 2.

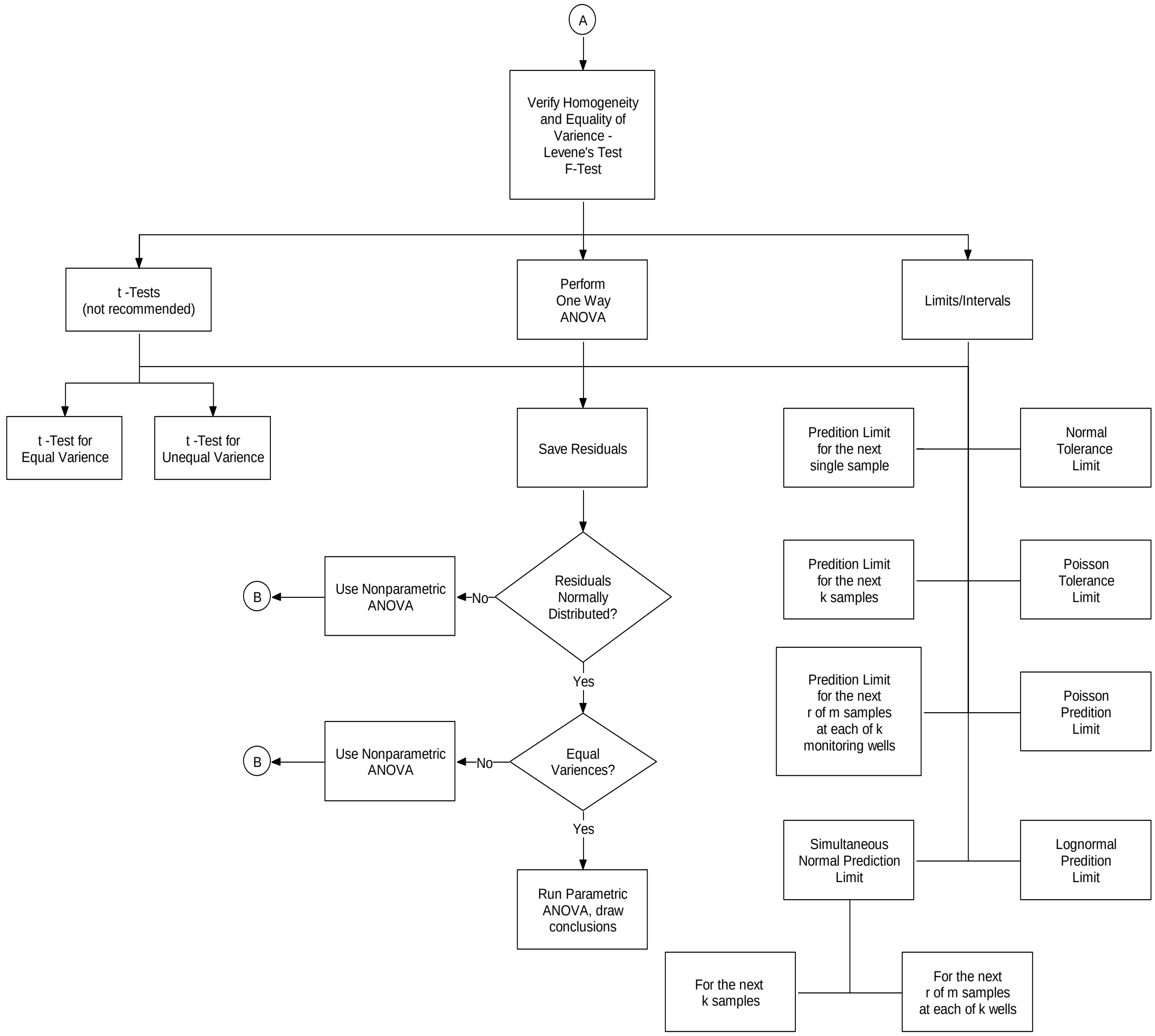

A
Verify Homogeneity and Equality of Varience - Levene's Test F-Test
t -Tests (not recommended)
Perform One Way ANOVA
Limits/Intervals
t -Test for Equal Varience
t -Test for Unequal Varience
Save Residuals
Residuals Normally Distributed?
No
Use Nonparametric ANOVA
B
Yes
Equal Variences?
No
Use Nonparametric ANOVA
B
Yes
Run Parametric ANOVA, draw conclusions
Predition Limit for the next single sample
Normal Tolerance Limit
Predition Limit for the next k samples
Poisson Tolerance Limit
Predition Limit for the next r of m samples at each of k monitoring wells
Poisson Predition Limit
Simultaneous Normal Prediction Limit
Lognormal Predition Limit
For the next k samples
For the next r of m samples at each of k wells

Figure 4-4. Statistical Decision Tree with Options for Groundwater Monitoring-Part 3.

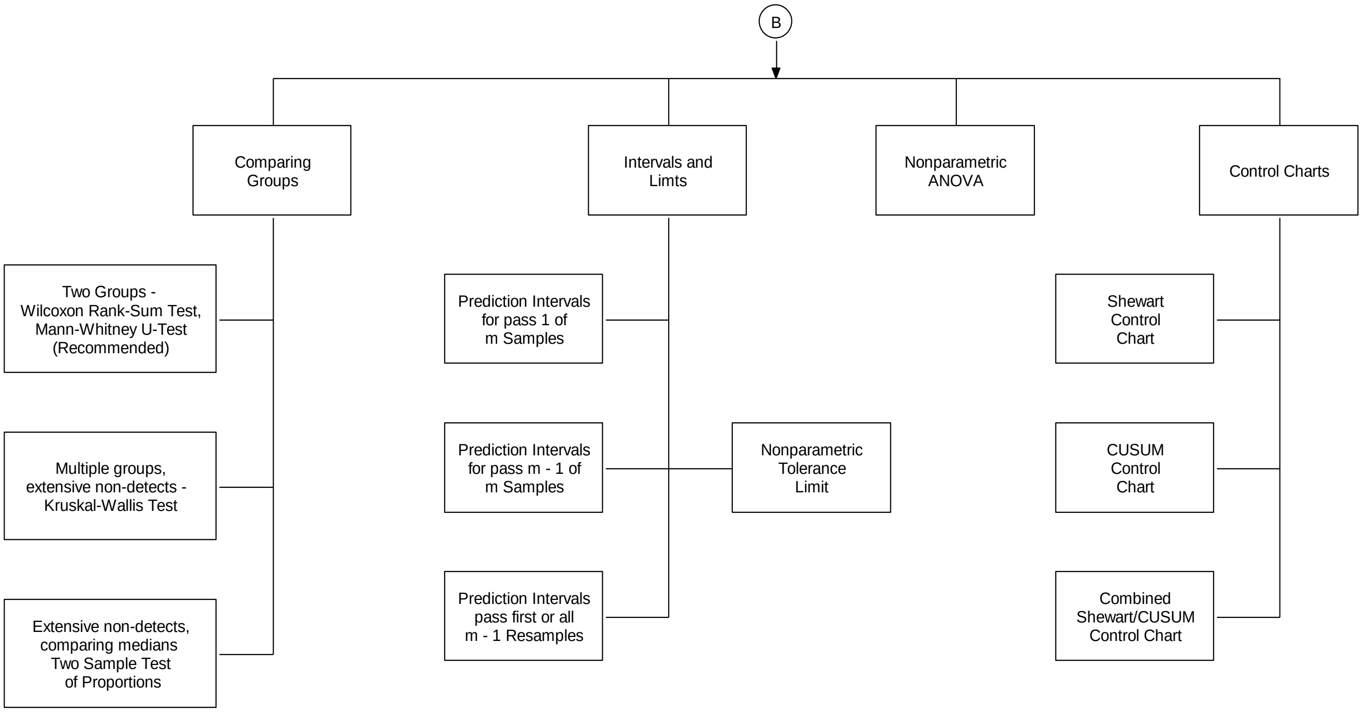

B
Comparing Groups
Intervals and Limts
Nonparametric ANOVA
Control Charts
Two Groups - Wilcoxon Rank-Sum Test, Mann-Whitney U-Test (Recommended)
Multiple groups, extensive non-detects - Kruskal-Wallis Test
Extensive non-detects, comparing medians Two Sample Test of Proportions
Prediction Intervals for pass 1 of m Samples
Prediction Intervals for pass m - 1 of m Samples
Prediction Intervals pass first or all m - 1 Resamples
Nonparametric Tolerance Limit
Shewart Control Chart
CUSUM Control Chart
Combined Shewart/CUSUM Control Chart

Figure 4-5. Statistical Decision Tree with Options for Groundwater Monitoring-Part 4.

4.6. Case Study 2—Shewart/CUSUM Monitoring. A groundwater plume at a site is currently being addressed via pumping and treating large amounts of groundwater. The system is very costly, and the site owner wishes to change the system configuration. Project regulators want to know whether changing the system (in this case, shutting off the treatment system) will increase measured trichloroethene (TCE) values near the leading edge of the plume. A special type of compliance monitoring was initiated to determine whether concentrations after system shutdown exceeded a “trigger” level. Table 4-2 lists the eight most recent TCE measurements at monitoring well B-37 prior to altering the system.

4.6.1. The sample mean for these data ($\bar{x}$) is 4.3 parts per billion (ppb) and the sample standard deviation (*s*) is 1.1 ppb. These values are used in statistical tests for normality, which did not indicate the data set is non-normal. (A hypothesis of normality cannot be rejected at the 90% significance level using any of the Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov, or D’Agostino tests [See Appendix F].)

4.6.2. Table 4-3 lists the measured TCE concentrations in this well over eight monitoring periods after system shutdown in mid-December 2002, and the associated Shewart/CUSUM statistical parameters (see Appendix Q). The Shewart/CUSUM calculations shown in the table are plotted in the Figure 4-6.

4.6.3. The quantities z_i and S_i (discussed in Appendix Q) were calculated to determine whether changing the system configuration resulted in an unacceptable change (i.e., increase) in the TCE concentration in Well B-37.

Table 4-2.
Eight Most Recent TCE Measurements in Well B-37

Well ID	Sample Date	TCEConcentration (µg/L)
B-37	7-Jun-99	3.0
B-37	29-Nov-99	3.2
B-37	26-Jun-00	4.5
B-37	3-Jan-01	5.8
B-37	16-May-01	5.9
B-37	4-Oct-01	3.2
B-37	27-Mar-02	4.6
B-37	10-Dec-02	4.3

4.6.4. The first out-of-control event occurred in winter 2003 when the z_i of 4.8 exceeded the Shewart threshold of 4.5. In addition, although the normalized concentration z_i decreases after the fifth sampling event following the start of shutdown, S_i continues to increase beyond and remains greater than the threshold of 5.0 for this quantity through fall 2004.

Table 4-3.
TCE Measurements and Shewart/CUSUM Calculations

Hypothetical Sampling Event	Sampling Period, *i*	TCE Concentration (µg/L)	z_i	$z_{i\text{-}1}$	S_i
Winter 2002	1	4.9	0.6	–0.4	0
Spring 2003	2	5.7	1.2	0.2	0.2
Summer 2003	3	6.0	1.4	0.4	0.7
Fall 2003	4	3.9	–0.4	–1.4	0.0
Winter 2003	5	9.8	4.8	3.8	3.8
Spring 2004	6	8.1	3.3	2.3	6.1
Summer 2004	7	7.5	2.8	1.8	8.0
Fall 2004	8	10.6	5.5	4.5	12.5

$\boldsymbol{z_i}$ = standardized result (or normalized concentration)
$\boldsymbol{S}_i$ = cumulative sum

4.6.5. The results of the testing showed that reconfiguring the system appeared to change the concentrations of TCE in this downgradient well at a statistically significant level. The reconfiguration was abandoned, and project planners began to reevaluate their understanding of groundwater movement at the site.

4.6.6. The Shewart/CUSUM method is commonly applied to landfills for detection monitoring, although it has obvious additional uses in other long-term monitoring applications. For instance, by looking for an insignificant change over time, a site stakeholder could suggest that monitoring at a natural attenuation site could be discontinued.

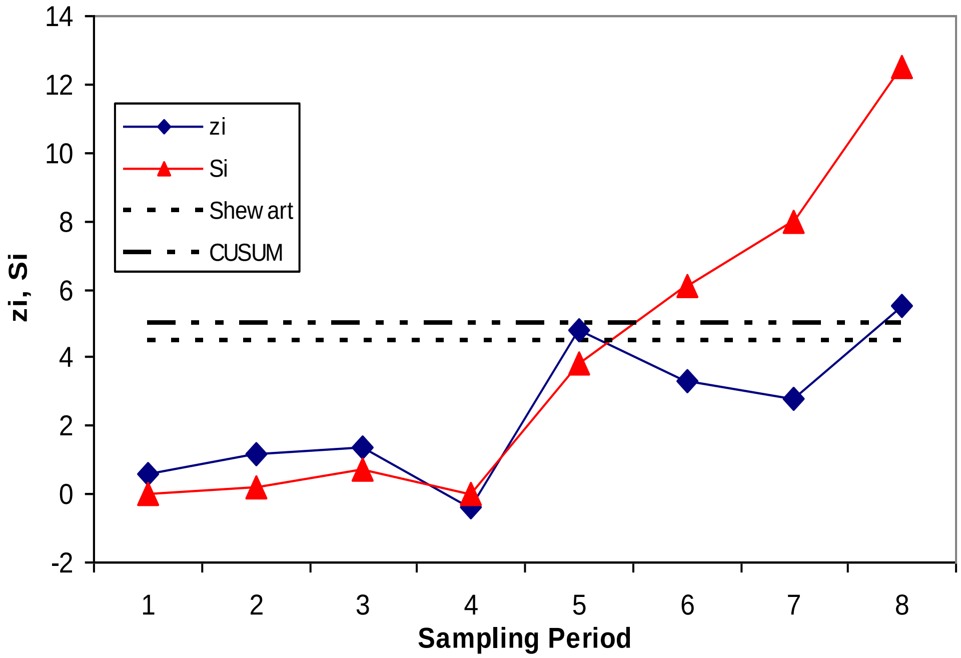


Figure 4-6. Shewart/CUSUM control chart, Well B-37.

4.7. Optimization**.** The process of optimization is similar in many ways to the process of sensitivity analysis. In both cases, one makes planned adjustments to the system and looks for changes in the outcome. The process of optimization involves assessing whether or not a change made in the system results in a beneficial outcome—improving system performance, for example, by reducing cost, increasing efficiency, or shortening the time to completion. This can be accomplished by comparing data taken after the adjustments have been made to historical data for the process using a variety of hypothesis testing tools.

4.7.1. It is also possible to examine trends in the system after taking into account seasonal and other forms of cyclic correlation. For example, when a time plot is examined for trend after a system modification, one may find that the slope of the time plot line changes, indicating a change in system performance. A time series plot is a graph showing how a parameter (e.g., TCE concentration) changes over time. A trend is a statistically significant change upward or downward with a certain degree of confidence. Whether or not that change is significant and an assessment of the magnitude of its impact can be addressed using trend tests such as Mann-Kendall and Sen's Slope Estimator.*

4.7.2. Another example of system optimization is in addressing such issues as the monitored analyte list and the frequency of sampling, both of which have economic implications and can have regulatory implications as well. As a hypothetical extreme case for illustration, assume that a monitoring well network must be sampled four times each year; that there are 10 wells in the network; and that each well is monitored for 50 constituents, all of which must be non-detects.

4.7.3. The statistics underlying the determination of a detection limit (e.g., if normality is assumed and the detection limit is the "Type I detection limit" or "critical value" in Appendix G) are such that there is only a 1% probability of a false positive at the detection limit while, as the statistics employed are one-sided, there is a 50% probability of a false negative at the detection limit. Thus, in the course of a given year, based on probability alone, the facility could falsely report itself in violation an average of 20 times, while falsely reporting compliance 1000 times (on the average). In fact, it can be demonstrated that simply because of the inherent Type I error rate associated with any statistical test, where literally thousands of such comparisons may be required, whether at the detection limit or otherwise, the probability of a false conclusion of violation approaches unity. Thus, it is always in the best interest of the regulated facility to limit the number of analytes for which one tests to the smallest possible number. Every permit renewal period or 5-year review should be used as an opportunity to further limit the analyte list. Even hypothetically, one can see that this approach is inefficient (costly), and reaching the goal of all non-detect is an example of a poorly defined quality objective. Detection limits can differ across laboratories and over time, and, clearly, they are not related to risk management in any way.

4.7.4. Another approach currently under study is the use of statistics to establish predictable correlation between the analyte of interest and some parameter that is more readily or cost-effectively measured than the analyte of interest. This "harbinger" or "calibration" approach has its roots in the commonly accepted practice of monitoring for indicator parameters such as pH, conductivity, total organic carbon, and total organic halides in place of specific analytes. If a rigorous regression analysis of historical data suggests a quantitative linkage between the concentration of arsenic and magnesium at a given site, it should be possible to delete, or at least reduce the frequency of analysis, for one or the other analyte, particularly in the case where both analytes have historically displayed compliant behavior. It would also be useful in this type of situation if a functional relationship and the uncertainty associated with that relationship could be established.

* Appendix P.

4.7.5. To assess the viability of monitored natural attenuation as a remedial alternative, it is essential to demonstrate: i) degradation of VOCs from parent products through to mineralization; and ii) correlation between that degradation and appropriate geochemical conditions. An example of assessing the correlation of parameters at a site in Maryland is illustrated in Case Study 3. Correlation measures show how strongly variables (or parameters) are related, or change with each other.

4.8. Case Study 3—Trend Analysis and Correlation in Natural Attenuation Data.

4.8.1. The data used for a site in Maryland were organized along a single geographic line, from the suspected source to a groundwater discharge zone located along a creek bed. Location was displayed in feet from the center of the suspected source. The parent constituent was PCE. The primary geochemical indicators of interest (for purposes of this case study) were dissolved oxygen (DO) and oxidation-reduction potential (redox).

Table 4-4.
Attenuation Data

Distance from Source (feet)	PCE (µg/L)	DO (mg/L)	Redox (mV)
0	320	0	–210
50	1430	0	–220
100	960	0.2	–170
150	780	0.3	–140
200	570	0.6	–80
250	630	0.5	–30
300	580	0.8	10
350	340	1.1	40
400	430	1.4	70
450	130	1.7	90
500	12	3.5	120

4.8.2. The data for the three parameters of interest are presented in Table 4-4. The data were then plotted against distance from the origin (source) to identify trends over distance. A Mann-Kendall trend analysis showed that PCE concentration decreased over distance. Redox and DO are positively correlated to one another with a Pearson's *r* value of 0.84. Geochemical understanding of natural attenuation requires that redox and DO should be inversely correlated to PCE concentration, and the Pearson's *r* values for DO and redox are –0.71 and –0.74, respectively. The results are displayed in the Figures 4-7 and 4-8. In summary, the results suggest that conditions for natural attenuation are present.

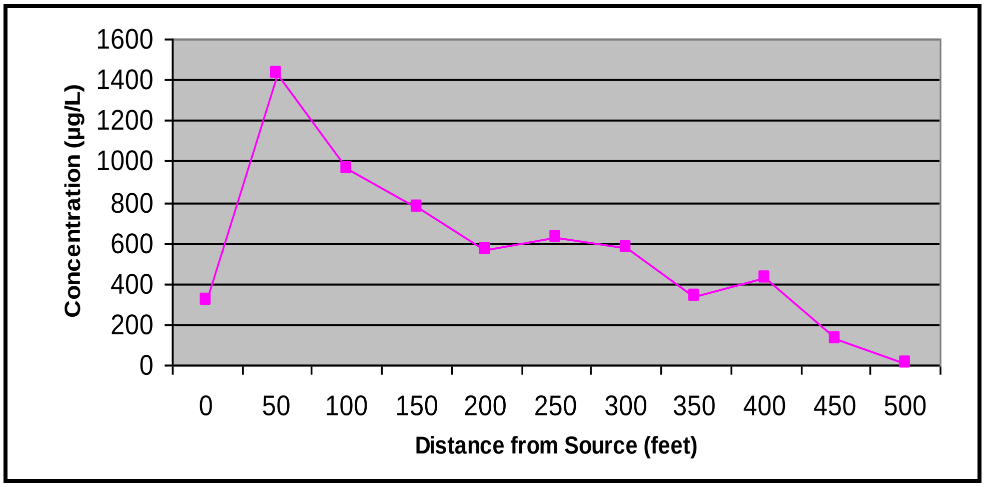


Figure 4-7. PCE Concentration Versus Distance.

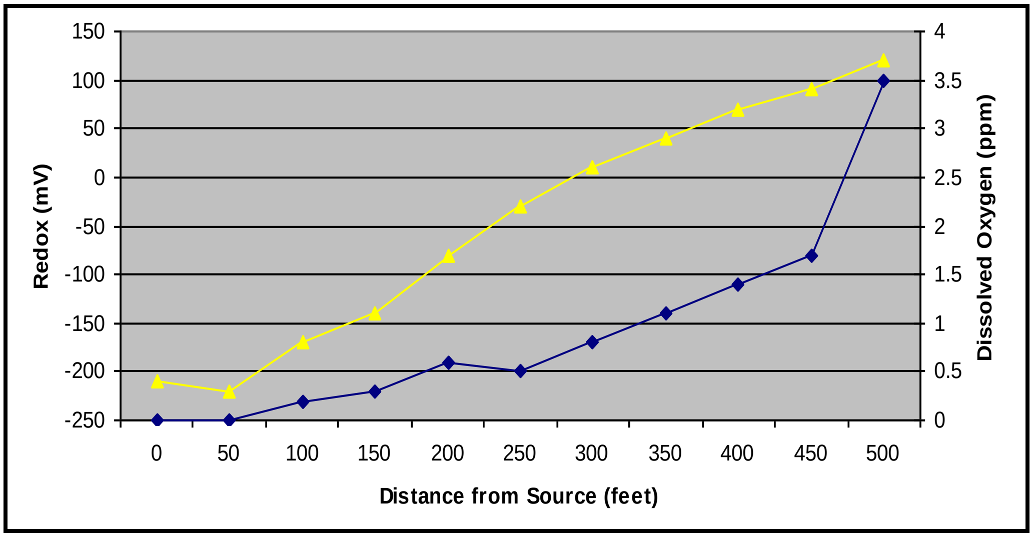


Figure 4-8. Geotechnical Parameters Versus Distance From Source.
(yellow triangle-redox; blue diamond-dissolved oxygen.

Section II
Applying Cleanup Levels

4.9. Introduction**.** When derived in accordance with USEPA's risk assessment guidance, risk-based cleanup levels are intended to represent the average contaminant concentration within the exposure unit that can be left on the site following remediation (Schulz and Griffin, 2001). In contrast, a "not-to-exceed" cleanup level drives remediation solutions that involve treating or removing any and all media with contaminant concentrations that exceed the cleanup level. The result is that applying a not-to-exceed level may result in over-remediation.

4.9.1. Calculated using risk assessment principles, the cleanup goal concentration is usually defined as an exposure unit concentration that will meet the target risk level agreed to by the design team and regulatory authorities. Some sample concentrations exceeding the cleanup objective can remain in place as long as the overall exposure concentration, calculated to a predetermined level of certainty, meets the cleanup goal (and likewise the agreed upon risk level). Because of the uncertainty associated with estimating the true average concentration of a contaminant at a site, USEPA recommends use of the 95% one-sided, upper confidence limit of the arithmetic mean (95% UCL) of the sample data to represent the exposure unit concentration term in risk assessments (EPA 9285.7-09A and EPA OSWER 9285.6-10). Consequently, a risk-based cleanup level should generally be interpreted as the 95% UCL of the contaminant concentration within the exposure unit following remediation.

4.9.2. However, draft USEPA guidance suggests specific situations in which application of the cleanup level as an area average may not be appropriate (USEPA, 2002) These include the following.

4.9.2.1. Exposure within the exposure unit is not random.

4.9.2.2. The cleanup level is based on acute rather than chronic exposure.

4.9.2.3. The cleanup level is not risk-based (i.e., it considers factors other than risk).

4.9.2.4. The quality of site characterization data is not optimal but it is not worth investing in additional sampling.

4.9.2.5. Given the site conditions (complexity, size, characterization, contaminant distribution), it is not cost-effective to do the necessary sampling and statistical analysis.

4.9.2.6. The community will not accept leaving soil with contaminant concentrations that exceed the cleanup level on the site.

4.9.3. If applying cleanup levels as an area average is appropriate, there are two basic approaches: i) using non-spatial statistical methods to determine a not-to-exceed concentration, and ii) using spatial statistical methods to iteratively re-calculate the UCL until the optimal "design line" for the remedial action is determined.

4.10. Determining Not-to-Exceed Concentrations Using Non-Spatial Statistics. Draft USEPA guidance (USEPA, 2002) defines the remedial action level (RAL) as the maximum concentration that may be left in place within an exposure unit such that the average concentration (or 95% UCL) within the exposure unit is at or below the cleanup level. Non-spatial techniques may be appropriate for calculating the RAL when there is no spatial correlation between contaminant concentrations, such as at a dump site where small, randomly located spots of high contaminant concentrations are interspersed with areas of lower concentrations. Non-spatial techniques are based on the mean and standard deviation of the sample contaminant concentration data and on how those metrics change as soils with high contaminant concentrations are replaced with post-remediation concentrations during remediation. The draft guidance describes two non-spatial statistical methods for calculating remedial action levels that ensure that post-remediation area average contaminant concentrations achieve cleanup levels: i) iterative truncation method, and ii) confidence response goal method. These methods are also reviewed in Schulz and Griffin (2001). Both methods can be applied in a spreadsheet calculation or programming language.

4.10.1. Iterative Truncation Method.

4.10.1.1. The iterative truncation method is based on the identifying and removing (truncating) high values in the sample concentration measurements (hot spots), replacing them with the post-remediation concentration (e.g., concentration in clean fill), and calculating the hypothetical post-remediation average concentration (95% UCL) in the exposure unit. Starting with the highest concentration in the data set, the process is repeated iteratively until the post-remediation 95% UCL is less than or equal to the cleanup level. The highest sample concentration remaining in the data set is designated the RAL.

4.10.1.2. This method is sensitive to the completeness of site characterization and the range of resultant sample concentrations. According to the draft USEPA guidance, to use this method with confidence, good site characterization through extensive, unbiased sampling is required and the resulting data must adequately represent random, long-term exposure to receptors. This method is not reliable when samples are not independently and randomly located.

4.10.2. Confidence Response Goal Method. Bowers et al. (1996) developed a method for calculating a confidence response goal (CRG), which, like the RAL, is a not-to-exceed level. This method can be applied at sites where there is a non-spatial, lognormal distribution of contamination (USEPA, 2002).

4.10.2.1. As described in the draft USEPA guidance, the basic premise of the method is that the CRG can be expressed as a function of the geometric mean and the geometric standard deviation of contaminant concentrations, and the desired reduction in exposure, which is defined as the ratio of average post-remediation concentration to the average pre-remediation concentration. The guidance provides a summary of the method, documents the equation for calculating the CRG, and refers the reader to the original paper (Bowers et al., 1996) for details on the derivation of the function.

4.10.2.2. The Schulz and Griffin (2001) review of the two non-spatial methods concludes that the CRG method is less sensitive than the iterative truncation method to changes in the highest sample concentrations and recommends the use of the CRG method when the contaminant distribution is lognormal.

4.10.3. Using Spatial Statistical Methods to Determine "Design Line" for Remediation. The distribution of contaminant concentrations may be spatially correlated at many sites where there is an original source or release that is subject to environmental fate and transport mechanisms. Contaminant concentrations in and around the original source or release may be higher than those at greater distances, or they may be higher where there is a mechanism of accumulation or an environmental "sink." Biased sampling is frequently applied in such cases because a high number of samples is desired in areas with high variance and uncertainty (for example, near the source area), and a lower number of samples is often sufficient to characterize areas with expected low variance and uncertainty. The concept of taking "step out" samples in the vicinity of sample locations where high contaminant concentrations are detected also introduces bias into the sampling plan. Geostatistical techniques are statistical procedures designed to process spatially correlated data (see Appendix R on Geostatistics). Unlike the non-spatial techniques, geostatistical techniques are well suited for evaluation of biased data sets.

4.10.3.1. The draft USEPA guidance presents an example of the determination of RALs using geostatistical techniques. The example has two simplifying features that can be found on many (but not all) sites: i) contamination that is surface only, and ii) the importance of a residential scenario. For this example, the steps for determining RALs are as follows.

4.10.3.1.1. Create an iso-concentration map of the site by modeling the spatial correlation underlying measured values.

4.10.3.1.2. Superimpose a grid of exposure units over the site and compute average contaminant concentrations in each exposure unit.

4.10.3.1.3. Identify zones that must be remediated to reduce average concentrations in all exposure units to the appropriate cleanup level. This is an iterative process, where the higher contaminant concentrations are replaced with post-remediation concentrations and average contaminant concentrations in each exposure unit are re-calculated. The final cutoff concentration is the RAL.

4.10.3.1.4. Use the original iso-concentration map to define zones with concentrations in excess of the RAL. The contoured zone is the area that requires remediation.

4.10.3.2. The draft guidance cautions against using geostatistical techniques if contaminant concentrations show a random, non-spatial pattern, or if the anticipated benefits from geostatistical analysis do not justify the costs. For example, even in cases of conservatively biased data, spatial statistical methods may not be warranted when non-spatial methods are determined to result in cleanup objectives that are both sufficiently conservative from the risk perspective and acceptable from the cleanup cost perspective. Additionally, conservatively biased, non-spatial methods may be needed from a practical view when adequate technical or computational resources are not available. Proponents of geostatistical techniques counter that presenting the site contamination and remediation results as spatial is a highly intuitive and visually powerful approach, and therefore enhances communication among the parties during risk management discussions. Available computational tools make it possible to find the point of diminishing returns where an increase in remediation has little effect on reducing risk in a cost-effective manner.

APPENDIX A

References

A-1. Required References.

ER 200-1-5
Policy for Implementation and Integrated Application of the U.S. Army Corps of Engineers Environmental Operating Principles and Doctrine

A-2. Related References.

A-2.1. Government Publications.

A-2.1.1. Army.

ER 5-1-11
U.S. Army Corps of Engineers Business Process.

EM 200-1-2
Technical Project Planning (TPP) Process,

EM 200-1-4, Volume I
Risk Assessment Handbook: Volume I - Human Health Evaluation.

EM 200-1-4, Volume II
Risk Assessment Handbook: Volume II - Environmental Evaluation.

EM 200-1-6
Environmental Quality—Chemical Quality Assurance for Hazardous, Toxic and Radioactive Waste (HTRW) Projects.

EM 200-1-10
Guidance for Evaluating Performance-Based Chemical Data.

EM 1110-1-502
Technical Guidelines for Hazardous and Toxic Waste Treatment and Cleanup Activities.

A-2.1.2. Navy.

UG-2049-ENV
U.S. Navy, Guidance for Environmental Background Analysis—Volume 1: Soil. DON, Naval Facilities Engineering Command.

A-2.1.3. Environmental Protection Agency.

EPA 68-W0-0025
Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities: Addendum to Interim Final Guidance. USEPA, Office of Solid Waste Management Division, June 1992.

EPA 230-R-92-14
Methods for Evaluating the Attainment of Cleanup Standards, Vol. 2: Ground Water, USEPA, Office of Policy, Planning, and Evaluation, Washington, D.C., 1992.

EPA 230-R-94-004
Statistical Methods for Evaluating the Attainment of Cleanup Standards, Vol. 3: Reference-Based Standards for Soils and Solid Media, USEPA, Office of Policy, Planning, and Evaluation, Washington, D.C., 1994.

EPA 230-R-95-005
EPA Observational Economy Series, Volume 1: Composite Sampling, Policy, Planning and Evaluation, August 1995.

EPA 230-R-95-006
EPA Observational Economy Series, Volume 2: Rank Set Sampling, Policy Planning and Evaluation, August 1995.

EPA 530/R-09-007
Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities Unified Guidance, EPA Office of Resource Conservation and Recovery, March 2009.

EPA 540-R-01-003
Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites. OSWER 9285.7-41, September 2002.

EPA 540-R-01-007
Comprehensive Five-Year Review Guidance, Office of Emergency and Remedial Response, Washington, D.C., OSWER 9355.7-03B-P, June 2001.

EPA 540-R-01-008
USEPA National Functional Guidelines for Inorganic Data Review, USEPA, Office of Superfund Remediation and Technology Innovation, Washington, D.C., July 2002.

EPA 540-R-02-002
Risk Assessment Guidance for Superfund: Volume III – Part A, Process for Conducting Probabilistic Risk Assessment. Office of Emergency and Remedial Response, Washington D.C., OSWER 9285.7-45, December 2001.

EPA 540/R-95/140
Superfund Program, Representative Sampling Guidance, Volume 2: Air (Short-Term Monitoring), Interim Final, Office of Solid Waste and Emergency Response, Washington, D.C., OSWER 9360.4-09, December 1995.

EPA 540/R-95/141
Superfund Program, Representative Sampling Guidance, Volume 1: Soil. Interim Final, Office of Solid Waste and Emergency Response, Washington, D.C., OSWER 9360.4-10, December 1995.

EPA 540-R-97-006
Ecological Risk Assessment Guidance for Superfund: Process Designing and Conducting Ecological Risk Assessments, Interim Final, USEPA, Solid Waste and Emergency Response, Washington, D.C., OSWER 9285.7-25, June 1997.

EPA 540-R-97-028
Superfund Method for the Determination of Releasable Asbestos in Soils and Bulk Materials, USEPA, Office of Solid Waste and Emergency Response, Washington, D.C., 1997.

EPA/540/S-96/500
Determination of Background Concentrations of Inorganics in Soils and Sediments at Hazardous Waste Sites by Breckenridge, R. P. and A. B. Crockett (U.S. Department of Energy). USEPA, Office of Solid Waste and Emergency Response, Washington, D.C. December, 1995.

EPA 600/4-82-029
Handbook for Sampling and Sample Preservation of Water and Wastewater. Environmental Monitoring and Support Laboratory, USEPA, Cincinnati, Ohio, 1982.

EPA 600/4-88/033
GEO-EAS. USEPA Environmental Monitoring Systems Laboratory, prepared by E. Englund and A. Sparks, 1988.

EPA 600/4-88-/040
Evaluation of Control Chart Methodologies for RCRA Waste Sites, Office of Research and Development, USEPA, Las Vegas, NV, 1989.

EPA/600/P-95/002
Exposure Factors Handbook. Office of Research and Development, National Center for Environmental Assessment, Washington D.C., August 1997.

EPA 600/R-97/006
Singh, A. K., A. Singh, and M. Engelhardt, The Lognormal Distribution in Environmental Applications, Technology Support Center Issue, Office of Research and Development, Office of Solid Waste and Emergency Response, December 1997.

EPA 903-F-94-001
EPA Region III, Hazardous Waste Management Division, Office of Superfund Programs, Philadelphia, PA, February 1994.

EPA 9285.7-09A
Guidance for Data Usability and Risk Assessment, Part A [Final]. USEPA, Office of Emergency and Remedial Response, Washington, D.C., April 1992.

EPA OB92-963373
Supplemental Guidance to RAGS: Calculating the Concentration Term. USEPA, Office of Solid Waste and Emergency Response, Washington, D.C., May 1992.

EPA QA/G-1
Guidance for Developing Quality Systems for Environmental Programs, EPA 240R-02/2008, USEPA, Office of Environmental Information, Washington, D.C., November 2002.

EPA QA/G-4
Guidance for the Data Quality Objectives Process, USEPA, Office of Environmental Information, EPA 600/R-96/055, QA/G-4, August 2000.

EPA QA/G-4D
Data Quality Objectives Decision Error Feasibility Trials Software(DEFT)—User’s Guide, USEPA, Office of Environmental Information, Washington, D.C, EPA/240/B-01/007, September 2001.

EPA QA/G-5
Guidance for Quality Assurance Project Plans, USEPA, Office of Environmental Information, Washington D.C., EPA/240/R-02/009, December 2002.

EPA QA/G-5S
Guidance on Choosing a Sampling Design for Environmental Data Collection, USEPA, Office of Environmental Information, Washington D.C., EPA/240/R-02/005, December 2002.

EPA QA/G-6
Guidance for Preparing Standard Operating Procedures, USEPA, Office of Environmental Information, Washington D.C., EPA/240/B-01/004, March 2001.

EPA QA/G-7
Guidance on Technical Audits and Related Assessments for Environmental Data Operations, USEPA, Office of Environmental Information, Washington D.C., EPA/600/R-99/080, January 2000.

EPA QA/G-8
Guidance on Environmental Data Verification and Data Validation, USEPA, Office of Environmental Information, Washington D.C., EPA/240/R-02/004, November 2002.

EPA QA/G-9R
Data Quality Assessment: A Reviewer’s Guide, USEPA, Office of Environmental Information, Washington, D.C. EPA/240/B-06/002, February 2006.

EPA QA/G-9S
Data Quality Assessment: Statistical Methods for Practitioners, USEPA, Office of Environmental Information, Washington, D.C., EPA/240/B-O6-003, February 2006.

EPA QA/G-10
Guidance for Developing a Training Program for Quality Systems, USEPA, Office of Environmental Information, Washington D.C., EPA/240/B-00/004, December 2000.

EPA QA/G-11
Guidance on Quality Assurance for Environmental Technology Design, Construction and Operation, USEPA, Office of Environmental Information, Washington D.C., EPA/240/B-05/001, January 2005.

EPA SOW No. 788
Contract Laboratory Program Statement of Work for Inorganics Analysis: Multi-media, Multi-concentration. Office of Emergency and Remedial Response, 1988.

EPA SW-846
On-line documentation for Test Methods for Evaluating Solid Wastes: Physical/Chemical Methods, 3rd Edition, Volume 2, Part III, Chapter 9 Sampling Plan, Rev. 0, SW-846, September 1986.

EPA (2002)
Guidance on Surface Soil Cleanup at Superfund Sites: Applying Cleanup Levels, Draft, Prepared by Industrial Economics, Incorporated for EPA, Office of Emergency and Remedial Response, Washington, D.C., January 2002.

OSWER 92-856-03
EPA, Exposure Factors Handbook, 1991.

OSWER 9285.7-41/EPA 540-R-01-003
EPA, Guidance for Comparing Background Chemical Concentrations in Soil at CERCLA, USEPA, Office of Emergency and Remedial Response, September 2002.

OSWER 9285.6-10
EPA, Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites, Office of Emergency and Remedial Response, Washington, D.C., OSWER 9285.6-10, December 2002.

OSWER 9360.4-16
Superfund Program, Representative Sampling Guidance, Volume 5: Water and Sediment, Part II – Groundwater, Interim Final, USEPA, Office of Solid Waste and Emergency Response, Washington, D.C. December 1995.

A-2.2 Non-government Publications.

ANSI/ASQC E4-1994
Specifications and Guidelines for Quality Systems for Environmental Data Collection and Environmental Technology Programs, ANSI/ASQC, Quality Press, 1994.

ASTM D-4210-89
American Society for Testing and Materials, Standard Practice for Intralaboratory Quality Control Procedures and a Discussion on Reporting Low-Level Data, 1996.

ASTM D5549-94e1
American Society for Testing Materials, Standard Guide for the Contents of Geostatistical Site Investigation Report, March 2000

Bowers et al. (1996)
Bowers, T. S., N. S. Shifrin, and B. L. Murphy, 1996. "Statistical Approach to Meeting Soil Cleanup Goals," Environmental Science & Technology, 30(5): 1437–1444, 1996.

Bohonak (2004)
Bohonak, A. J., RMA: Software for Reduced Major Axis Regression. 3 September 2004. http:/www.bio.sdsu.edu/pub/andy/rma.html. v.1.17.

Carmer and Swanson (1973)
Carmer, S. G. and M. R. Swanson, "Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods," Journal of the American Statistical Association, 68(314), 1973.

Clark (1979)
Clark, I., Practical Geostatistics, Applied Science Publishers, London, 1979.

Conover (1999)
Conover, W. J., Practical Nonparametric Statistics, John Wiley & Sons, NY, 1999.

Cressie (1991)
Cressie, N., Statistics for spatial data, revised edition, John Wiley & Sons, NY, 1993.

Currie (1968)
Currie, L. A., "Limits for Qualitative Detection and Quantitative Determination: Application to Radiochemistry," Analytical Chemistry 40, 586-593, March 1968.

Devore (1987)
Devore, J. L., Probability and Statistics for Engineering and the Sciences. Brooks/Cole Publishing Company, 1987.

Georgian and Osborn (2003)
Georgian, T., and K. Osborn, Quality Assurance: Good Practice, Regulation, and Law, Volume 10, Number 1, Taylor and Francis, Inc. Philadelphia, PA, 2003.

Gibbons and Coleman (2001)
Gibbons, R. D., and D. E. Coleman, Statistical Methods for Detection and Quantification of Environmental Contamination, Wiley-Interscience, July 2001.

Gibbons (1994)
Gibbons, R. D., Statistical Methods for Groundwater Monitoring, John Wiley & Sons, Inc., 1994.

Gilbert (1987)
Gilbert, R. O., Statistical Methods for Environmental Pollution Monitoring. John Wiley & Sons, Inc., 1987.

Gnanadesikan (1997)
Gnanadesikan, R., Methods for Statistical Data Analysis of Multivariate Observations, 2nd Edition, John Wiley & Sons, Inc., New York, 1997.

Goovaerts (1997)
Goovaerts P., Geostatistics for natural resource evaluation, Oxford University Press, New York, 1997.

Hahn, 1970
Hahn, G. J., "Statistical Intervals for a Normal Population, Part II. Formulas, Assumptions, Some Derivations," Journal of Quality Technology, Vol. 2, No.4, 195 – 206, October 1970.

Hahn and Meeker (1991)
Hahn, G. J., and W. Q. Meeker, Statistical Intervals: A Guide for Practitioners. John Wiley & Sons, Inc., 1991.

Hall et al. (1975)
Hall, I. J., R. R. Prairie, and C. K. Motlagh, "Non-Parametric Prediction Intervals," Journal of Quality Technology, 7(3), 1975.

Helsel and Hirsch (1992)
Helsel, D. R., and R. M. Hirsch, Studies in Environmental Science 49—Statistical Methods in Water Resources, Amsterdam, Elsevier, 1992.

Helsel and Hirsch (2003)
Helsel, D. R., and R. M. Hirsch, Statistical Methods in Water Resources. U.S. Geological Survey, Techniques of Water-Resources Investigations, Book 4, Chapter A3. http://water.usgs.gov/pubs/twri/twri4a3/html/pdf_new.html, May 2003.

Helsel (2005)
Helsel, D. R., Non-detects and Data Analysis: Statistics for Censored Environmental Data, John Wiley & Sons, N.J., 2005.

Hoaglin et al. (1983)
Hoaglin, D. C., F. Mosteller, and J. W. Tukey, Understanding Robust and Exploratory Data Analysis, John Wiley and Sons, Inc., 1983.

Hockman and Lucas (1987)
Hockman, K. K., and J. M. Lucas, “Variability Reduction Through Sub-vessel CUMSUM Control,” Journal of Quality Technology, Vol. 19, pp. 113-121, 1987.

Kvanli et al. (1996)
Kvanli, A.H., C.S. Guynes, and R.J. Pavur, Introduction to Business Statistics: A Computer Integrated, Data Analysis Approach, West Publishing Company, 1996.

Lehmann (1975)
Lehmann, E. L., Nonparametrics: Statistical Methods Based on Ranks, Holden-Day, Inc., 1975.

Lucas (1982)
Lucas, J. M. “Combined Shewhart-CUMSUM Quality Control Schemes,” Journal of Quality Technology, 14: 51–59, 1982.

Mason et al. (1989)
Mason, R L., R. F. Gunst, and J. L. Hess, Statistical Design and Analysis of Experiments: With Applications to Engineering and Science, John Wiley & Sons, Inc., 1989.

Meyers (1997)
Meyers, J. C., Geostatistical Error Management: Quantifying Uncertainty for Environmental Sampling and Mapping, Van Nostrand Reinhold, New York, 1997.

Milton and Arnold (1990)
Milton, J. S., and J. C. Arnold, Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences, McGraw-Hill, Inc., 1990.

Montgomery (1997)
Montgomery D. C., Design and Analysis of Experiments, 4th Edition, John Wiley & Sons, 1997.

Moore (1999)
Moore, D. S., and G. P. McCabe. Introduction to the Practice of Statistics, 3rd Edition, W.H. Freeman and Company, New York, 1999.

Moser (2000)
Moser, D., Risk Analysis Program Developments. Risk Analysis For Water Resources Investments Newsletter. United States Army Corps of Engineers. Summer 2000, Issue 4.

Noether (1987)
Noether, G. E., “Sample Size Determination for Some Common Nonparametric Tests,” Journal of the American Statistical Association, 82(398), Theory and Methods, 1987.

Schulz and Griffin (2001)
Schulz, T. W., and S. Griffin, “Practical methods for meeting remediation goals at hazardous waste sites,” Risk Analysis, 21(1): 43–52.

Snedecor and Cochran (1982)
Snedecor, G. W., and W. G. Cochran, Statistical Methods, 7th Edition, Iowa State University Press, 1982.

Warton (2005)
Warton, David I., “Bivariate line fitting methods for allometry,” March 17, 2005, preprint, http://www.ncbi.nlm.nih.gov/pubmed/16573844

APPENDIX B

Statistical Tables

Table B-1.
Binomial Distribution

n	*k*	*p*=0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
1	0	0.9500	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000	0.4500	0.4000	0.3500	0.3000	0.2500	0.2000	0.1500	0.1000	0.05000
	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500	0.2025	0.1600	0.1225	0.09000	0.06250	0.04000	0.02250	0.01000	0.002500
	1	0.9975	0.9900	0.9775	0.9600	0.9375	0.9100	0.8775	0.8400	0.7975	0.7500	0.6975	0.6400	0.5775	0.5100	0.4375	0.3600	0.2775	0.1900	0.09750
	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0	0.8574	0.729	0.6141	0.512	0.4219	0.343	0.2746	0.216	0.1664	0.125	0.09113	0.064	0.04288	0.027	0.01563	8.000E-03	3.375E-03	1.000E-03	1.250E-04
	1	0.9928	0.972	0.9393	0.896	0.8438	0.784	0.7183	0.648	0.5748	0.5	0.4253	0.352	0.2818	0.216	0.1563	0.104	0.06075	0.028	7.250E-03
	2	0.9999	0.999	0.9966	0.992	0.9844	0.973	0.9571	0.936	0.9089	0.875	0.8336	0.784	0.7254	0.657	0.5781	0.488	0.3859	0.271	0.1426
	3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	0	0.8145	0.6561	0.522	0.4096	0.3164	0.2401	0.1785	0.1296	0.09151	0.0625	0.04101	0.0256	0.01501	8.100E-03	3.906E-03	1.600E-03	5.062E-04	1.000E-04	6.250E-06
	1	0.986	0.9477	0.8905	0.8192	0.7383	0.6517	0.563	0.4752	0.391	0.3125	0.2415	0.1792	0.1265	0.0837	0.05078	0.0272	0.01198	3.700E-03	4.812E-04
	2	0.9995	0.9963	0.988	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875	0.609	0.5248	0.437	0.3483	0.2617	0.1808	0.1095	0.0523	0.01402
	3	1.000	0.9999	0.9995	0.9984	0.9961	0.9919	0.985	0.9744	0.959	0.9375	0.9085	0.8704	0.8215	0.7599	0.6836	0.5904	0.478	0.3439	0.1855
	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.116	0.07776	0.05033	0.03125	0.01845	0.01024	5.252E-03	2.430E-03	9.766E-04	3.200E-04	7.594E-05	1.000E-05	3.125E-07
	1	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.337	0.2562	0.1875	0.1312	0.08704	0.05402	0.03078	0.01562	6.720E-03	2.227E-03	4.600E-04	3.000E-05
	2	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5	0.4069	0.3174	0.2352	0.1631	0.1035	0.05792	0.02661	8.560E-03	1.158E-03
	3	1.000	0.9995	0.9978	0.9933	0.9844	0.9692	0.946	0.913	0.8688	0.8125	0.7438	0.663	0.5716	0.4718	0.3672	0.2627	0.1648	0.08146	0.02259
	4	1.000	1.000	0.9999	0.9997	0.999	0.9976	0.9947	0.9898	0.9815	0.9688	0.9497	0.9222	0.884	0.8319	0.7627	0.6723	0.5563	0.4095	0.2262

n	*k*	*p*=0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
	5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6	**0**	0.7351	0.5314	0.3771	0.2621	0.178	0.1176	0.07542	0.04666	0.02768	0.01563	8.304E-03	4.096E-03	1.838E-03	7.290E-04	2.441E-04	6.400E-05	1.139E-05	1.000E-06	1.562E-08
	1	0.9672	0.8857	0.7765	0.6554	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094	0.0692	0.04096	0.02232	0.01094	4.639E-03	1.600E-03	3.987E-04	5.500E-05	1.797E-06
	2	0.9978	0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3438	0.2553	0.1792	0.1174	0.07047	0.0376	0.01696	5.885E-03	1.270E-03	8.641E-05
	3	0.9999	0.9987	0.9941	0.983	0.9624	0.9295	0.8826	0.8208	0.7447	0.6563	0.5585	0.4557	0.3529	0.2557	0.1694	0.09888	0.04734	0.01585	2.230E-03
	4	1.000	0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.959	0.9308	0.8906	0.8364	0.7667	0.6809	0.5798	0.4661	0.3446	0.2235	0.1143	0.03277
	5	1.000	1.000	1.000	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844	0.9723	0.9533	0.9246	0.8824	0.822	0.7379	0.6229	0.4686	0.2649
	6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7	**0**	0.6983	0.4783	0.3206	0.2097	0.1335	0.08235	0.04902	0.02799	0.01522	7.813E-03	3.737E-03	1.638E-03	6.434E-04	2.187E-04	6.104E-05	1.280E-05	1.709E-06	1.000E-07	7.812E-10
	1	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625	0.03571	0.01884	9.008E-03	3.791E-03	1.343E-03	3.712E-04	6.948E-05	6.400E-06	1.047E-07
	2	0.9962	0.9743	0.9262	0.852	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266	0.1529	0.09626	0.05561	0.0288	0.01288	4.672E-03	1.222E-03	1.765E-04	6.027E-06
	3	0.9998	0.9973	0.9879	0.9667	0.9294	0.874	0.8002	0.7102	0.6083	0.5	0.3917	0.2898	0.1998	0.126	0.07056	0.03334	0.0121	2.728E-03	1.936E-04
	4	1.000	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734	0.6836	0.5801	0.4677	0.3529	0.2436	0.148	0.07377	0.02569	3.757E-03
	5	1.000	1.000	0.9999	0.9996	0.9987	0.9962	0.991	0.9812	0.9643	0.9375	0.8976	0.8414	0.7662	0.6706	0.5551	0.4233	0.2834	0.1497	0.04438
	6	1.000	1.000	1.000	1.000	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922	0.9848	0.972	0.951	0.9176	0.8665	0.7903	0.6794	0.5217	0.3017
	7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8	**0**	0.6634	0.4305	0.2725	0.1678	0.1001	0.05765	0.03186	0.0168	8.373E-03	3.906E-03	1.682E-03	6.554E-04	2.252E-04	6.561E-05	1.526E-05	2.560E-06	2.563E-07	1.000E-08	3.906E-11
	1	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.06318	0.03516	0.01812	8.520E-03	3.571E-03	1.290E-03	3.815E-04	8.448E-05	1.187E-05	7.300E-07	5.977E-09
	2	0.9942	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445	0.08846	0.04981	0.02532	0.01129	4.227E-03	1.231E-03	2.423E-04	2.341E-05	4.008E-07
	3	0.9996	0.995	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.477	0.3633	0.2604	0.1737	0.1061	0.05797	0.0273	0.01041	2.854E-03	4.316E-04	1.540E-05
	4	1.000	0.9996	0.9971	0.9896	0.9727	0.942	0.8939	0.8263	0.7396	0.6367	0.523	0.4059	0.2936	0.1941	0.1138	0.05628	0.02135	5.024E-03	3.718E-04
	5	1.000	1.000	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555	0.7799	0.6846	0.5722	0.4482	0.3215	0.2031	0.1052	3.809E-02	5.788E-03
	6	1.000	1.000	1.000	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648	0.9368	0.8936	0.8309	0.7447	0.6329	0.4967	0.3428	0.1869	0.05724
	7	1.000	1.000	1.000	1.000	1.000	0.9999	0.9998	0.9993	0.9983	0.9961	0.9916	0.9832	0.9681	0.9424	0.8999	0.8322	0.7275	0.5695	0.3366
	8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
9	**0**	0.6302	0.3874	0.2316	0.1342	0.07508	0.04035	0.02071	0.01008	4.605E-03	1.953E-03	7.567E-04	2.621E-04	7.882E-05	1.968E-05	3.815E-06	5.120E-07	3.844E-08	1.000E-09	1.953E-12
	1	0.9288	0.7748	0.5995	0.4362	0.3003	0.196	0.1211	0.07054	0.03852	0.01953	9.080E-03	3.801E-03	1.396E-03	4.330E-04	1.068E-04	1.894E-05	1.999E-06	8.200E-08	3.359E-10

n	***k***	***p*=0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
	2	0.9916	0.947	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.08984	0.04977	0.02503	0.01118	4.291E-03	1.343E-03	3.139E-04	4.644E-05	2.998E-06	2.572E-08
	3	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539	0.1658	0.09935	0.05359	0.02529	9.995E-03	3.066E-03	6.340E-04	6.423E-05	1.151E-06
	4	1.000	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5	0.3786	0.2666	0.1717	0.09881	0.04893	0.01958	5.629E-03	8.909E-04	3.322E-05
	5	1.000	0.9999	0.9994	0.9969	0.99	0.9747	0.9464	0.9006	0.8342	0.7461	0.6386	0.5174	0.3911	0.2703	0.1657	0.08564	0.03393	8.331E-03	6.426E-04
	6	1.000	1.000	1.000	0.9997	0.9987	0.9957	0.9888	0.975	0.9502	0.9102	0.8505	0.7682	0.6627	0.5372	0.3993	0.2618	0.1409	0.05297	8.361E-03
	7	1.000	1.000	1.000	1.000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805	0.9615	0.9295	0.8789	0.804	0.6997	0.5638	0.4005	0.2252	0.07121
	8	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.9992	0.998	0.9954	0.9899	0.9793	0.9596	0.9249	0.8658	0.7684	0.6126	0.3698
	9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
10	**0**	0.5987	0.3487	0.1969	0.1074	0.05631	0.02825	0.01346	6.047E-03	2.533E-03	9.766E-04	3.405E-04	1.049E-04	2.759E-05	5.905E-06	9.537E-07	1.024E-07	5.767E-09	1.000E-10	9.766E-14
	1	0.9139	0.7361	0.5443	0.3758	0.244	0.1493	0.08595	0.04636	0.02326	0.01074	4.502E-03	1.678E-03	5.399E-04	1.437E-04	2.956E-05	4.198E-06	3.325E-07	9.100E-09	1.865E-11
	2	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.09956	0.05469	0.02739	0.01229	4.821E-03	1.590E-03	4.158E-04	7.793E-05	8.665E-06	3.736E-07	1.605E-09
	3	0.999	0.9872	0.95	0.8791	0.7759	0.6496	0.5138	0.3823	0.266	0.1719	0.102	0.05476	0.02602	0.01059	3.506E-03	8.644E-04	1.346E-04	9.122E-06	8.198E-08
	4	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.377	0.2616	0.1662	0.09493	0.04735	0.01973	6.369E-03	1.383E-03	1.469E-04	2.755E-06
	5	1.000	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.623	0.4956	0.3669	0.2485	0.1503	0.07813	0.03279	9.874E-03	1.635E-03	6.369E-05
	6	1.000	1.000	0.9999	0.9991	0.9965	0.9894	0.974	0.9452	0.898	0.8281	0.734	0.6177	0.4862	0.3504	0.2241	0.1209	0.04997	0.0128	1.028E-03
	7	1.000	1.000	1.000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453	0.9004	0.8327	0.7384	0.6172	0.4744	0.3222	0.1798	0.07019	0.0115
	8	1.000	1.000	1.000	1.000	1.000	0.9999	0.9995	0.9983	0.9955	0.9893	0.9767	0.9536	0.914	0.8507	0.756	0.6242	0.4557	0.2639	0.08614
	9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.999	0.9975	0.994	0.9865	0.9718	0.9437	0.8926	0.8031	0.6513	0.4013
	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
11	**0**	0.5688	0.3138	0.1673	0.0859	0.04224	0.01977	8.7510E-03	3.6280E-03	1.3930E-03	4.8830E-04	1.532E-04	4.194E-05	9.655E-06	1.771E-06	2.384E-07	2.048E-08	8.650E-10	1.000E-11	4.883E-15
	1	0.8981	0.6974	0.4922	0.3221	0.1971	0.113	0.06058	0.03023	0.01393	5.8590E-03	2.213E-03	7.340E-04	2.069E-04	4.724E-05	8.106E-06	9.216E-07	5.478E-08	1.000E-09	1.025E-12
	2	0.9848	0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.06522	3.2710E-02	0.0148	5.924E-03	2.038E-03	5.777E-04	1.261E-04	1.894E-05	1.582E-06	4.555E-08	9.797E-11
	3	0.9984	0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133	0.06096	0.02928	0.01224	4.291E-03	1.188E-03	2.352E-04	2.755E-05	1.248E-06	5.624E-09
	4	0.9999	0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5328	0.3971	0.2744	0.1738	0.09935	0.05014	0.02162	7.561E-03	1.965E-03	3.219E-04	2.290E-05	2.156E-07
	5	1.000	0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7535	0.6331	0.5	0.3669	0.2465	0.1487	0.07822	0.03433	0.01165	2.657E-03	2.957E-04	5.801E-06
	6	1.000	1.000	0.9997	0.998	0.9924	0.9784	0.9499	0.9006	0.8262	0.7256	0.6029	0.4672	0.3317	0.2103	0.1146	0.05041	0.01589	2.751E-03	1.119E-04
	7	1.000	1.000	1.000	0.9998	0.9988	0.9957	0.9878	0.9707	0.939	0.8867	0.8089	0.7037	0.5744	0.4304	0.2867	0.1611	0.06944	0.01853	1.552E-03
	8	1.000	1.000	1.000	1.000	0.9999	0.9994	0.998	0.9941	0.9852	0.9673	0.9348	0.8811	0.7999	0.6873	0.5448	0.3826	0.2212	0.08956	0.01524

n	*k*	*p*=0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
	9	1.000	1.000	1.000	1.000	1.000	1.000	0.9998	0.9993	0.9978	0.9941	0.9861	0.9698	0.9394	0.887	0.8029	0.6779	0.5078	0.3026	0.1019
	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9998	0.9995	0.9986	0.9964	0.9912	0.9802	0.9578	0.9141	0.8327	0.6862	0.4312
	11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
12	**0**	0.5404	0.2824	0.1422	0.06872	0.03168	0.01384	5.688E-03	2.177E-03	7.662E-04	0.0002441	6.895E-05	1.678E-05	3.379E-06	5.314E-07	5.960E-08	4.096E-09	1.297E-10	1.000E-12	2.441E-16
	1	0.8816	0.659	0.4435	0.2749	0.1584	0.08503	0.04244	0.01959	8.289E-03	0.003174	1.080E-03	3.188E-04	7.869E-05	1.541E-05	2.205E-06	2.007E-07	8.952E-09	1.090E-10	5.591E-14
	2	0.9804	0.8891	0.7358	0.5583	0.3907	0.2528	0.1513	0.08344	0.04214	0.01929	7.878E-03	2.810E-03	8.479E-04	2.064E-04	3.761E-05	4.526E-06	2.839E-07	5.455E-09	5.873E-12
	3	0.9978	0.9744	0.9078	0.7946	0.6488	0.4925	0.3467	0.2253	0.1345	0.073	0.03557	0.01527	5.610E-03	1.692E-03	3.917E-04	6.220E-05	5.478E-06	1.658E-07	3.743E-10
	4	0.9998	0.9957	0.9761	0.9274	0.8424	0.7237	0.5833	0.4382	0.3044	0.1938	0.1117	0.05731	0.02551	9.489E-03	2.782E-03	5.812E-04	7.170E-05	3.414E-06	1.612E-08
	5	1.000	0.9995	0.9954	0.9806	0.9456	0.8822	0.7873	0.6652	0.5269	0.3872	0.2607	0.1582	0.08463	0.0386	0.01425	3.903E-03	6.721E-04	5.018E-05	4.949E-07
	6	1.000	0.9999	0.9993	0.9961	0.9857	0.9614	0.9154	0.8418	0.7393	0.6128	0.4731	0.3348	0.2127	0.1178	0.0544	0.01941	4.642E-03	5.412E-04	1.111E-05
	7	1.000	1.000	0.9999	0.9994	0.9972	0.9905	0.9745	0.9427	0.8883	0.8062	0.6956	0.5618	0.4167	0.2763	0.1576	0.07256	0.02392	4.329E-03	1.839E-04
	8	1.000	1.000	1.000	0.9999	0.9996	0.9983	0.9944	0.9847	0.9644	0.927	0.8655	0.7747	0.6533	0.5075	0.3512	0.2054	0.09221	0.02564	2.236E-03
	9	1.000	1.000	1.000	1.000	1.000	0.9998	0.9992	0.9972	0.9921	0.9807	0.9579	0.9166	0.8487	0.7472	0.6093	0.4417	0.2642	0.1109	0.01957
	10	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.9989	0.9968	0.9917	0.9804	0.9576	0.915	0.8416	0.7251	0.5565	0.341	0.1184
	11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9998	0.9992	0.9978	0.9943	0.9862	0.9683	0.9313	0.8578	0.7176	0.4596
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
13	**0**	0.5133	0.2542	0.1209	0.05498	0.02376	9.689E-03	3.697E-03	1.306E-03	4.214E-04	1.221E-04	3.103E-05	6.711E-06	1.183E-06	1.594E-07	1.490E-08	8.192E-10	1.946E-11	1.000E-13	1.221E-17
	1	0.8646	0.6213	0.3983	0.2336	0.1267	0.06367	0.02958	0.01263	4.904E-03	1.709E-03	5.240E-04	1.376E-04	2.974E-05	4.996E-06	5.960E-07	4.342E-08	1.453E-09	1.180E-11	3.027E-15
	2	0.9755	0.8661	0.692	0.5017	0.3326	0.2025	0.1132	0.0579	0.02691	0.01123	4.139E-03	1.315E-03	3.479E-04	7.270E-05	1.106E-05	1.066E-06	5.020E-08	6.436E-10	3.468E-13
	3	0.9969	0.9658	0.882	0.7473	0.5843	0.4206	0.2783	0.1686	0.09292	0.04614	0.02034	7.793E-03	2.515E-03	6.520E-04	1.261E-04	1.606E-05	1.063E-06	2.149E-08	2.429E-11
	4	0.9997	0.9935	0.9658	0.9009	0.794	0.6543	0.5005	0.353	0.2279	0.1334	0.06985	0.03208	0.01257	4.031E-03	9.891E-04	1.660E-04	1.541E-05	4.906E-07	1.162E-09
	5	1.000	0.9991	0.9925	0.97	0.9198	0.8346	0.7159	0.5744	0.4268	0.2905	0.1788	0.09767	0.0462	0.01822	5.649E-03	1.246E-03	1.618E-04	8.090E-06	4.006E-08
	6	1.000	0.9999	0.9987	0.993	0.9757	0.9376	0.8705	0.7712	0.6437	0.5	0.3563	0.2288	0.1295	0.06238	0.02429	7.004E-03	1.268E-03	9.929E-05	1.026E-06
	7	1.000	1.000	0.9998	0.9988	0.9944	0.9818	0.9538	0.9023	0.8212	0.7095	0.5732	0.4256	0.2841	0.1654	0.08021	0.03004	7.534E-03	9.200E-04	1.975E-05
	8	1.000	1.000	1.000	0.9998	0.999	0.996	0.9874	0.9679	0.9302	0.8666	0.7721	0.647	0.4995	0.3457	0.206	0.09913	0.03416	6.460E-03	2.866E-04
	9	1.000	1.000	1.000	1.000	0.9999	0.9993	0.9975	0.9922	0.9797	0.9539	0.9071	0.8314	0.7217	0.5794	0.4157	0.2527	0.118	0.03416	3.103E-03
	10	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.9987	0.9959	0.9888	0.9731	0.9421	0.8868	0.7975	0.6674	0.4983	0.308	0.1339	0.02451
	11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9995	0.9983	0.9951	0.9874	0.9704	0.9363	0.8733	0.7664	0.6017	0.3787	0.1354
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9996	0.9987	0.9963	0.9903	0.9762	0.945	0.8791	0.7458	0.4867

n	*k*	*p*=0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
	13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
14	**0**	0.4877	0.2288	0.1028	0.04398	0.01782	6.782E-03	2.403E-03	7.836E-04	2.318E-04	6.104E-05	1.396E-05	2.684E-06	4.140E-07	4.783E-08	3.725E-09	1.638E-10	2.919E-12	1.000E-14	6.104E-19
	1	0.847	0.5846	0.3567	0.1979	0.101	0.04748	0.02052	8.098E-03	2.887E-03	9.155E-04	2.529E-04	5.906E-05	1.118E-05	1.610E-06	1.602E-07	9.339E-09	2.345E-10	1.270E-12	1.630E-16
	2	0.9699	0.8416	0.6479	0.4481	0.2811	0.1608	0.08393	0.03979	0.01701	6.470E-03	2.151E-03	6.087E-04	1.411E-04	2.531E-05	3.211E-06	2.479E-07	8.765E-09	7.498E-11	2.021E-14
	3	0.9958	0.9559	0.8535	0.6982	0.5213	0.3552	0.2205	0.1243	0.06322	0.02869	0.01143	3.906E-03	1.106E-03	2.465E-04	3.982E-05	4.065E-06	2.021E-07	2.729E-09	1.544E-12
	4	0.9996	0.9908	0.9533	0.8702	0.7415	0.5842	0.4227	0.2793	0.1672	0.08978	0.04262	0.01751	6.035E-03	1.666E-03	3.419E-04	4.605E-05	3.215E-06	6.840E-08	8.117E-11
	5	1.000	0.9985	0.9885	0.9561	0.8883	0.7805	0.6405	0.4859	0.3373	0.212	0.1189	0.05832	0.02434	8.289E-03	2.154E-03	3.819E-04	3.736E-05	1.251E-06	3.107E-09
	6	1.000	0.9998	0.9978	0.9884	0.9617	0.9067	0.8164	0.6925	0.5461	0.3953	0.2586	0.1501	0.07534	0.03147	0.01031	2.397E-03	3.276E-04	1.721E-05	8.934E-08
	7	1.000	1.000	0.9997	0.9976	0.9897	0.9685	0.9247	0.8499	0.7414	0.6047	0.4539	0.3075	0.1836	0.09328	0.03827	0.01161	2.207E-03	1.814E-04	1.962E-06
	8	1.000	1.000	1.000	0.9996	0.9978	0.9917	0.9757	0.9417	0.8811	0.788	0.6627	0.5141	0.3595	0.2195	0.1117	0.04385	0.01153	1.474E-03	3.309E-05
	9	1.000	1.000	1.000	1.000	0.9997	0.9983	0.994	0.9825	0.9574	0.9102	0.8328	0.7207	0.5773	0.4158	0.2585	0.1298	0.04674	9.230E-03	4.274E-04
	10	1.000	1.000	1.000	1.000	1.000	0.9998	0.9989	0.9961	0.9886	0.9713	0.9368	0.8757	0.7795	0.6448	0.4787	0.3018	0.1465	0.04413	4.173E-03
	11	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9994	0.9978	0.9935	0.983	0.9602	0.9161	0.8392	0.7189	0.5519	0.3521	0.1584	0.03005
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.9991	0.9971	0.9919	0.9795	0.9525	0.899	0.8021	0.6433	0.4154	0.153
	13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9998	0.9992	0.9976	0.9932	0.9822	0.956	0.8972	0.7712	0.5123
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
15	**0**	0.4633	0.2059	0.08735	0.03518	0.01336	4.748E-03	1.562E-03	4.702E-04	1.275E-04	3.052E-05	6.283E-06	1.074E-06	1.449E-07	1.435E-08	9.313E-10	3.277E-11	4.379E-13	1.000E-15	3.052E-20
	1	0.829	0.549	0.3186	0.1671	0.08018	0.03527	0.01418	5.172E-03	1.692E-03	4.883E-04	1.215E-04	2.523E-05	4.181E-06	5.166E-07	4.284E-08	1.999E-09	3.766E-11	1.360E-13	8.728E-18
	2	0.9638	0.8159	0.6042	0.398	0.2361	0.1268	0.06173	0.02711	0.01065	3.693E-03	1.107E-03	2.789E-04	5.665E-05	8.719E-06	9.229E-07	5.705E-08	1.514E-09	8.641E-12	1.165E-15
	3	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.04242	0.01758	6.327E-03	1.928E-03	4.789E-04	9.166E-05	1.236E-05	1.011E-06	3.777E-08	3.403E-10	9.641E-14
	4	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.05923	0.02547	9.348E-03	2.831E-03	6.722E-04	1.153E-04	1.246E-05	6.541E-07	9.296E-09	5.525E-12
	5	0.9999	0.9978	0.9832	0.9389	0.8516	0.7216	0.5643	0.4032	0.2608	0.1509	0.07693	0.03383	0.01244	3.653E-03	7.949E-04	1.132E-04	8.338E-06	1.866E-07	2.324E-10
	6	1.000	0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036	0.1818	0.09505	0.04219	0.01524	4.193E-03	7.850E-04	8.090E-05	2.846E-06	7.418E-09
	7	1.000	1.000	0.9994	0.9958	0.9827	0.95	0.8868	0.7869	0.6535	0.5	0.3465	0.2131	0.1132	0.05001	0.0173	4.240E-03	6.096E-04	3.362E-05	1.830E-07
	8	1.000	1.000	0.9999	0.9992	0.9958	0.9848	0.9578	0.905	0.8182	0.6964	0.5478	0.3902	0.2452	0.1311	0.05662	0.01806	3.606E-03	3.106E-04	3.518E-06
	9	1.000	1.000	1.000	0.9999	0.9992	0.9963	0.9876	0.9662	0.9231	0.8491	0.7392	0.5968	0.4357	0.2784	0.1484	0.06105	0.01681	2.250E-03	5.281E-05
	10	1.000	1.000	1.000	1.000	0.9999	0.9993	0.9972	0.9907	0.9745	0.9408	0.8796	0.7827	0.6481	0.4845	0.3135	0.1642	0.06171	0.01272	6.147E-04
	11	1.000	1.000	1.000	1.000	1.000	0.9999	0.9995	0.9981	0.9937	0.9824	0.9576	0.9095	0.8273	0.7031	0.5387	0.3518	0.1773	0.05556	5.467E-03

n	*k*	*p*=0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
	12	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.9989	0.9963	0.9893	0.9729	0.9383	0.8732	0.7639	0.602	0.3958	0.1841	0.0362
	13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9995	0.9983	0.9948	0.9858	0.9647	0.9198	0.8329	0.6814	0.451	0.171
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9995	0.9984	0.9953	0.9866	0.9648	0.9126	0.7941	0.5367
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
16	**0**	0.4401	0.1853	0.07425	0.02815	0.01002	3.323E-03	1.015E-03	2.821E-04	7.011E-05	1.526E-05	2.827E-06	4.295E-07	5.071E-08	4.305E-09	2.328E-10	6.554E-12	6.568E-14	1.000E-16	1.526E-21
	1	0.8108	0.5147	0.2839	0.1407	0.06348	0.02611	9.763E-03	3.291E-03	9.880E-04	2.594E-04	5.812E-05	1.074E-05	1.558E-06	1.650E-07	1.141E-08	4.260E-10	6.021E-12	1.450E-14	4.654E-19
	2	0.9571	0.7892	0.5614	0.3518	0.1971	0.09936	0.04509	0.01834	6.620E-03	2.090E-03	5.650E-04	1.267E-04	2.254E-05	2.977E-06	2.629E-07	1.301E-08	2.591E-10	9.865E-13	6.657E-17
	3	0.993	0.9316	0.7899	0.5981	0.405	0.2459	0.1339	0.06515	0.02813	0.01064	3.456E-03	9.385E-04	2.044E-04	3.360E-05	3.783E-06	2.479E-07	6.952E-09	4.181E-11	5.928E-15
	4	0.9991	0.983	0.9209	0.7982	0.6302	0.4499	0.2892	0.1666	0.08531	0.03841	0.01494	4.896E-03	1.302E-03	2.658E-04	3.811E-05	3.301E-06	1.302E-07	1.236E-09	3.678E-13
	5	0.9999	0.9967	0.9765	0.9183	0.8103	0.6598	0.49	0.3288	0.1976	0.1051	0.04862	0.01914	6.196E-03	1.566E-03	2.852E-04	3.261E-05	1.807E-06	2.703E-08	1.687E-11
	6	1.000	0.9995	0.9944	0.9733	0.9204	0.8247	0.6881	0.5272	0.366	0.2272	0.1241	0.05832	0.02286	7.130E-03	1.644E-03	2.476E-04	1.922E-05	4.526E-07	5.917E-10
	7	1.000	0.9999	0.9989	0.993	0.9729	0.9256	0.8406	0.7161	0.5629	0.4018	0.2559	0.1423	0.06706	0.02567	7.470E-03	1.476E-03	1.602E-04	5.924E-06	1.620E-08
	8	1.000	1.000	0.9998	0.9985	0.9925	0.9743	0.9329	0.8577	0.7441	0.5982	0.4371	0.2839	0.1594	0.07435	0.02713	7.004E-03	1.059E-03	6.133E-05	3.497E-07
	9	1.000	1.000	1.000	0.9998	0.9984	0.9929	0.9771	0.9417	0.8759	0.7728	0.634	0.4728	0.3119	0.1753	0.07956	0.02666	5.586E-03	5.045E-04	5.983E-06
	10	1.000	1.000	1.000	1.000	0.9997	0.9984	0.9938	0.9809	0.9514	0.8949	0.8024	0.6712	0.51	0.3402	0.1897	0.08169	0.02354	3.297E-03	8.090E-05
	11	1.000	1.000	1.000	1.000	1.000	0.9997	0.9987	0.9951	0.9851	0.9616	0.9147	0.8334	0.7108	0.5501	0.3698	0.2018	0.07905	0.017	8.573E-04
	12	1.000	1.000	1.000	1.000	1.000	1.000	0.9998	0.9991	0.9965	0.9894	0.9719	0.9349	0.8661	0.7541	0.595	0.4019	0.2101	0.06841	7.004E-03
	13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9994	0.9979	0.9934	0.9817	0.9549	0.9006	0.8029	0.6482	0.4386	0.2108	0.04294
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.999	0.9967	0.9902	0.9739	0.9365	0.8593	0.7161	0.4853	0.1892
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.999	0.9967	0.99	0.9719	0.9257	0.8147	0.5599
	16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
17	**0**	0.4181	0.1668	0.06311	0.02252	7.517E-03	2.326E-03	6.600E-04	1.693E-04	3.856E-05	7.629E-06	1.272E-06	1.718E-07	1.775E-08	1.291E-09	5.821E-11	1.311E-12	9.853E-15	1.000E-17	7.629E-23
	1	0.7922	0.4818	0.2525	0.1182	0.05011	0.01928	6.701E-03	2.088E-03	5.749E-04	1.373E-04	2.771E-05	4.553E-06	5.781E-07	5.252E-08	3.027E-09	9.044E-11	9.590E-13	1.540E-15	2.472E-20
	2	0.9497	0.7618	0.5198	0.3096	0.1637	0.07739	0.03273	0.01232	4.086E-03	1.175E-03	2.862E-04	5.712E-05	8.903E-06	1.009E-06	7.427E-08	2.943E-09	4.399E-11	1.117E-13	3.770E-18
	3	0.9912	0.9174	0.7556	0.5489	0.353	0.2019	0.1028	0.04642	0.01845	6.363E-03	1.866E-03	4.514E-04	8.621E-05	1.216E-05	1.143E-06	5.999E-08	1.263E-09	5.069E-12	3.596E-16
	4	0.9988	0.9779	0.9013	0.7582	0.5739	0.3887	0.2348	0.126	0.05958	0.02452	8.623E-03	2.521E-03	5.887E-04	1.033E-04	1.236E-05	8.586E-07	2.544E-08	1.612E-10	2.402E-14
	5	0.9999	0.9953	0.9681	0.8943	0.7653	0.5968	0.4197	0.2639	0.1471	0.07173	0.0301	0.01059	3.015E-03	6.560E-04	9.989E-05	9.164E-06	3.817E-07	3.815E-09	1.193E-12
	6	1.000	0.9992	0.9917	0.9623	0.8929	0.7752	0.6188	0.4478	0.2902	0.1662	0.08259	0.03481	0.01203	3.235E-03	6.250E-04	7.561E-05	4.419E-06	6.959E-08	4.561E-11
	7	1.000	0.9999	0.9983	0.9891	0.9598	0.8954	0.7872	0.6405	0.4743	0.3145	0.1834	0.0919	0.03833	0.01269	3.101E-03	4.932E-04	4.037E-05	9.998E-07	1.372E-09

n	*k*	*p*=0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
	8	1.000	1.000	0.9997	0.9974	0.9876	0.9597	0.9006	0.8011	0.6626	0.5	0.3374	0.1989	0.09938	0.04028	0.01238	2.581E-03	2.950E-04	1.146E-05	3.287E-08
	9	1.000	1.000	1.000	0.9995	0.9969	0.9873	0.9617	0.9081	0.8166	0.6855	0.5257	0.3595	0.2128	0.1046	0.04024	0.01093	1.738E-03	1.056E-04	6.314E-07
	10	1.000	1.000	1.000	0.9999	0.9994	0.9968	0.988	0.9652	0.9174	0.8338	0.7098	0.5522	0.3812	0.2248	0.1071	0.03766	8.280E-03	7.838E-04	9.728E-06
	11	1.000	1.000	1.000	1.000	0.9999	0.9993	0.997	0.9894	0.9699	0.9283	0.8529	0.7361	0.5803	0.4032	0.2347	0.1057	0.03187	4.667E-03	1.197E-04
	12	1.000	1.000	1.000	1.000	1.000	0.9999	0.9994	0.9975	0.9914	0.9755	0.9404	0.874	0.7652	0.6113	0.4261	0.2418	0.09871	0.02214	1.165E-03
	13	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9995	0.9981	0.9936	0.9816	0.9536	0.8972	0.7981	0.647	0.4511	0.2444	0.08264	8.801E-03
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.9988	0.9959	0.9877	0.9673	0.9226	0.8363	0.6904	0.4802	0.2382	0.05025
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9994	0.9979	0.9933	0.9807	0.9499	0.8818	0.7475	0.5182	0.2078
	16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9998	0.9993	0.9977	0.9925	0.9775	0.9369	0.8332	0.5819
	17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
18	**0**	0.3972	0.1501	0.05365	0.01801	5.638E-03	1.628E-03	4.290E-04	1.016E-04	2.121E-05	3.815E-06	5.726E-07	6.872E-08	6.212E-09	3.874E-10	1.455E-11	2.621E-13	1.478E-15	1.000E-18	3.815E-24
	1	0.7735	0.4503	0.2241	0.09908	0.03946	0.01419	4.587E-03	1.320E-03	3.336E-04	7.248E-05	1.317E-05	1.924E-06	2.139E-07	1.666E-08	8.004E-10	1.914E-11	1.522E-13	1.630E-16	1.308E-21
	2	0.9419	0.7338	0.4797	0.2713	0.1353	0.05995	0.02362	8.226E-03	2.506E-03	6.561E-04	1.440E-04	2.558E-05	3.492E-06	3.394E-07	2.084E-08	6.609E-10	7.413E-12	1.256E-14	2.120E-19
	3	0.9891	0.9018	0.7202	0.501	0.3057	0.1646	0.07827	0.03278	0.01198	3.769E-03	9.971E-04	2.148E-04	3.596E-05	4.355E-06	3.414E-07	1.435E-08	2.269E-10	6.074E-13	2.156E-17
	4	0.9985	0.9718	0.8794	0.7164	0.5187	0.3327	0.1886	0.09417	0.04107	0.01544	4.907E-03	1.279E-03	2.621E-04	3.950E-05	3.948E-06	2.197E-07	4.890E-09	2.068E-11	1.543E-15
	5	0.9998	0.9936	0.9581	0.8671	0.7175	0.5344	0.355	0.2088	0.1077	0.04813	0.01829	5.750E-03	1.438E-03	2.691E-04	3.425E-05	2.520E-06	7.888E-08	5.266E-10	8.247E-14
	6	1.000	0.9988	0.9882	0.9487	0.861	0.7217	0.5491	0.3743	0.2258	0.1189	0.05372	0.02028	6.169E-03	1.430E-03	2.312E-04	2.245E-05	9.873E-07	1.039E-08	3.414E-12
	7	1.000	0.9998	0.9973	0.9837	0.9431	0.8593	0.7283	0.5634	0.3915	0.2403	0.128	0.05765	0.02123	6.073E-03	1.244E-03	1.591E-04	9.812E-06	1.626E-07	1.119E-10
	8	1.000	1.000	0.9995	0.9957	0.9807	0.9404	0.8609	0.7368	0.5778	0.4073	0.2527	0.1347	0.05969	0.02097	5.422E-03	9.109E-04	7.857E-05	2.046E-06	2.947E-09
	9	1.000	1.000	0.9999	0.9991	0.9946	0.979	0.9403	0.8653	0.7473	0.5927	0.4222	0.2632	0.1391	0.05959	0.01935	4.252E-03	5.115E-04	2.088E-05	6.280E-08
	10	1.000	1.000	1.000	0.9998	0.9988	0.9939	0.9788	0.9424	0.872	0.7597	0.6085	0.4366	0.2717	0.1407	0.05695	0.01628	2.719E-03	1.735E-04	1.086E-06
	11	1.000	1.000	1.000	1.000	0.9998	0.9986	0.9938	0.9797	0.9463	0.8811	0.7742	0.6257	0.4509	0.2783	0.139	0.05127	0.01182	1.172E-03	1.523E-05
	12	1.000	1.000	1.000	1.000	1.000	0.9997	0.9986	0.9942	0.9817	0.9519	0.8923	0.7912	0.645	0.4656	0.2825	0.1329	0.0419	6.415E-03	1.720E-04
	13	1.000	1.000	1.000	1.000	1.000	1.000	0.9997	0.9987	0.9951	0.9846	0.9589	0.9058	0.8114	0.6673	0.4813	0.2836	0.1206	0.02819	1.546E-03
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9998	0.999	0.9962	0.988	0.9672	0.9217	0.8354	0.6943	0.499	0.2798	0.0982	0.01087
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9993	0.9975	0.9918	0.9764	0.94	0.8647	0.7287	0.5203	0.2662	0.05813
	16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.9987	0.9954	0.9858	0.9605	0.9009	0.7759	0.5497	0.2265
	17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9996	0.9984	0.9944	0.982	0.9464	0.8499	0.6028
	18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n	*k*	*p*=0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
19	**0**	0.3774	0.1351	0.0456	0.01441	4.228E-03	1.140E-03	2.788E-04	6.094E-05	1.167E-05	1.907E-06	2.577E-07	2.749E-08	2.174E-09	1.162E-10	3.638E-12	5.243E-14	2.217E-16	1.000E-19	1.907E-25
	1	0.7547	0.4203	0.1985	0.08287	0.03101	0.01042	3.132E-03	8.328E-04	1.930E-04	3.815E-05	6.241E-06	8.109E-07	7.889E-08	5.269E-09	2.110E-10	4.037E-12	2.409E-14	1.720E-17	6.905E-23
	2	0.9335	0.7054	0.4413	0.2369	0.1113	0.04622	0.01696	5.464E-03	1.528E-03	3.643E-04	7.206E-05	1.139E-05	1.361E-06	1.135E-07	5.810E-09	1.475E-10	1.241E-12	1.402E-15	1.184E-20
	3	0.9868	0.885	0.6841	0.4551	0.2631	0.1332	0.05914	0.02296	7.719E-03	2.213E-03	5.279E-04	1.013E-04	1.486E-05	1.544E-06	1.010E-07	3.399E-09	4.033E-11	7.204E-14	1.280E-18
	4	0.998	0.9648	0.8556	0.6733	0.4654	0.2822	0.15	0.06961	0.02798	9.605E-03	2.756E-03	6.407E-04	1.151E-04	1.490E-05	1.243E-06	5.542E-08	9.263E-10	2.615E-12	9.762E-17
	5	0.9998	0.9914	0.9463	0.8369	0.6678	0.4739	0.2968	0.1629	0.07771	0.03178	0.01093	3.068E-03	6.736E-04	1.084E-04	1.152E-05	6.797E-07	1.599E-08	7.128E-11	5.589E-15
	6	1.000	0.9983	0.9837	0.9324	0.8251	0.6655	0.4812	0.3081	0.1727	0.08353	0.03423	0.01156	3.094E-03	6.173E-04	8.348E-05	6.506E-06	2.151E-07	1.513E-09	2.491E-13
	7	1.000	0.9997	0.9959	0.9767	0.9225	0.818	0.6656	0.4878	0.3169	0.1796	0.08713	0.03523	0.01144	2.823E-03	4.844E-04	4.979E-05	2.311E-06	2.561E-08	8.840E-12
	8	1.000	1.000	0.9992	0.9933	0.9713	0.9161	0.8145	0.6675	0.494	0.3238	0.1841	0.08847	0.03469	0.01054	2.288E-03	3.095E-04	2.013E-05	3.510E-07	2.537E-10
	9	1.000	1.000	0.9999	0.9984	0.9911	0.9674	0.9125	0.8139	0.671	0.5	0.329	0.1861	0.08747	0.03255	8.903E-03	1.579E-03	1.435E-04	3.930E-06	5.939E-09
	10	1.000	1.000	1.000	0.9997	0.9977	0.9895	0.9653	0.9115	0.8159	0.6762	0.506	0.3325	0.1855	0.08392	0.02875	6.658E-03	8.427E-04	3.614E-05	1.140E-07
	11	1.000	1.000	1.000	1.000	0.9995	0.9972	0.9886	0.9648	0.9129	0.8204	0.6831	0.5122	0.3344	0.182	0.07746	0.02328	4.084E-03	2.733E-04	1.793E-06
	12	1.000	1.000	1.000	1.000	0.9999	0.9994	0.9969	0.9884	0.9658	0.9165	0.8273	0.6919	0.5188	0.3345	0.1749	0.0676	0.01633	1.696E-03	2.306E-05
	13	1.000	1.000	1.000	1.000	1.000	0.9999	0.9993	0.9969	0.9891	0.9682	0.9223	0.8371	0.7032	0.5261	0.3322	0.1631	0.0537	8.593E-03	2.407E-04
	14	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9994	0.9972	0.9904	0.972	0.9304	0.85	0.7178	0.5346	0.3267	0.1444	0.03519	2.013E-03
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9995	0.9978	0.9923	0.977	0.9409	0.8668	0.7369	0.5449	0.3159	0.115	0.01324
	16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9996	0.9985	0.9945	0.983	0.9538	0.8887	0.7631	0.5587	0.2946	0.06655
	17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9998	0.9992	0.9969	0.9896	0.969	0.9171	0.8015	0.5797	0.2453
	18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9997	0.9989	0.9958	0.9856	0.9544	0.8649	0.6226
	19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
20	**0**	0.3585	0.1216	0.03876	0.01153	3.171E-03	7.979E-04	1.812E-04	3.656E-05	6.416E-06	9.537E-07	1.159E-07	1.100E-08	7.610E-10	3.487E-11	9.095E-13	1.049E-14	3.325E-17	1.000E-20	9.537E-27
	1	0.7358	0.3917	0.1756	0.06918	0.02431	7.637E-03	2.133E-03	5.240E-04	1.114E-04	2.003E-05	2.950E-06	3.408E-07	2.903E-08	1.662E-09	5.548E-11	8.493E-13	3.802E-15	1.810E-18	3.633E-24
	2	0.9245	0.6769	0.4049	0.2061	0.09126	0.03548	0.01212	3.611E-03	9.274E-04	2.012E-04	3.586E-05	5.041E-06	5.277E-07	3.773E-08	1.611E-09	3.273E-11	2.067E-13	1.557E-16	6.578E-22
	3	0.9841	0.867	0.6477	0.4114	0.2252	0.1071	0.04438	0.01596	4.933E-03	1.288E-03	2.772E-04	4.734E-05	6.084E-06	5.427E-07	2.960E-08	7.978E-10	7.105E-12	8.466E-15	7.523E-20
	4	0.9974	0.9568	0.8298	0.6296	0.4148	0.2375	0.1182	0.05095	0.01886	5.909E-03	1.531E-03	3.170E-04	4.994E-05	5.550E-06	3.865E-07	1.380E-08	1.732E-10	3.263E-13	6.097E-18
	5	0.9997	0.9887	0.9327	0.8042	0.6172	0.4164	0.2454	0.1256	0.05533	0.02069	6.434E-03	1.612E-03	3.106E-04	4.294E-05	3.813E-06	1.803E-07	3.186E-09	9.481E-12	3.722E-16
	6	1.000	0.9976	0.9781	0.9133	0.7858	0.608	0.4166	0.25	0.1299	0.05766	0.02141	6.466E-03	1.521E-03	2.610E-04	2.951E-05	1.845E-06	4.586E-08	2.155E-10	1.776E-14
	7	1.000	0.9996	0.9941	0.9679	0.8982	0.7723	0.601	0.4159	0.252	0.1316	0.05803	0.02103	6.015E-03	1.279E-03	1.837E-04	1.516E-05	5.295E-07	3.923E-09	6.786E-13
	8	1.000	0.9999	0.9987	0.99	0.9591	0.8867	0.7624	0.5956	0.4143	0.2517	0.1308	0.05653	0.01958	5.138E-03	9.354E-04	1.017E-04	4.983E-06	5.815E-08	2.108E-11

n	*k*	*p*=0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
	9	1.000	1.000	0.9998	0.9974	0.9861	0.952	0.8782	0.7553	0.5914	0.4119	0.2493	0.1275	0.05317	0.01714	3.942E-03	5.634E-04	3.863E-05	7.089E-07	5.380E-10
	10	1.000	1.000	1.000	0.9994	0.9961	0.9829	0.9468	0.8725	0.7507	0.5881	0.4086	0.2447	0.1218	0.04796	0.01386	2.5950E-03	2.484E-04	7.151E-06	1.134E-08
	11	1.000	1.000	1.000	0.9999	0.9991	0.9949	0.9804	0.9435	0.8692	0.7483	0.5857	0.4044	0.2376	0.1133	0.04093	9.982E-03	1.329E-03	5.986E-05	1.979E-07
	12	1.000	1.000	1.000	1.000	0.9998	0.9987	0.994	0.979	0.942	0.8684	0.748	0.5841	0.399	0.2277	0.1018	0.03214	5.921E-03	4.156E-04	2.857E-06
	13	1.000	1.000	1.000	1.000	1.000	0.9997	0.9985	0.9935	0.9786	0.9423	0.8701	0.75	0.5834	0.392	0.2142	0.08669	0.02194	2.386E-03	3.395E-05
	14	1.000	1.000	1.000	1.000	1.000	1.000	0.9997	0.9984	0.9936	0.9793	0.9447	0.8744	0.7546	0.5836	0.3828	0.1958	0.06731	0.01125	3.293E-04
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9997	0.9985	0.9941	0.9811	0.949	0.8818	0.7625	0.5852	0.3704	0.1702	0.04317	2.574E-03
	16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9997	0.9987	0.9951	0.984	0.9556	0.8929	0.7748	0.5886	0.3523	0.133	0.0159
	17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9998	0.9991	0.9964	0.9879	0.9645	0.9087	0.7939	0.5951	0.3231	0.07548
	18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9999	0.9995	0.9979	0.9924	0.9757	0.9308	0.8244	0.6083	0.2642
	19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.9998	0.9992	0.9968	0.9885	0.9612	0.8784	0.6415
	20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

K has the binomial distribution with parameters *n* and *p*. The entries are the values of $P(K \leq k)$ for *p* ranging from 0.05 to 0.95 for values of *n* ranging from 1 to 20.

For $n > 20$, the *q*th quantile of *K* (a binomial random variable) may be approximated using the formula: $K_q = np + Z_q\,[np(1\text{-}p)]^{1/2}$, where Z_q is the q^{th} quantile of the standard normal distribution.

Table B-2.
Percentiles of the Chi-Square Distribution

	p									
df	**0.005**	**0.010**	**0.025**	**0.050**	**0.1**	**0.900**	**0.950**	**0.975**	**0.990**	**0.995**
1	3.93E-05	0.000157	0.000982	0.003932	0.01579	2.706	3.841	5.024	6.635	7.879
2	0.01003	0.0201	0.05064	0.1026	0.2107	4.605	5.991	7.378	9.21	10.6
3	0.07172	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.34	12.84
4	0.207	0.2971	0.4844	0.7107	1.064	7.779	9.488	11.14	13.28	14.86
5	0.4117	0.5543	0.8312	1.145	1.61	9.236	11.07	12.83	15.09	16.75
6	0.6757	0.8721	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55
7	0.9893	1.239	1.69	2.167	2.833	12.02	14.07	16.01	18.48	20.28
8	1.344	1.646	2.18	2.733	3.49	13.36	15.51	17.53	20.09	21.95
9	1.735	2.088	2.7	3.325	4.168	14.68	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.94	4.865	15.99	18.31	20.48	23.21	25.19
11	2.603	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72	26.76
12	3.074	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22	28.3
13	3.565	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69	29.82
14	4.075	4.66	5.629	6.571	7.79	21.06	23.68	26.12	29.14	31.32
15	4.601	5.229	6.262	7.261	8.547	22.31	25	27.49	30.58	32.8
16	5.142	5.812	6.908	7.962	9.312	23.54	26.3	28.85	32	34.27
17	5.697	6.408	7.564	8.672	10.09	24.77	27.59	30.19	33.41	35.72
18	6.265	7.015	8.231	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.844	7.633	8.907	10.12	11.65	27.2	30.14	32.85	36.19	38.58
20	7.434	8.26	9.591	10.85	12.44	28.41	31.41	34.17	37.57	40
21	8.034	8.897	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.4
22	8.643	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.8
23	9.26	10.2	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.886	10.86	12.4	13.85	15.66	33.2	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	12.2	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	20.6	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	67.5	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.4	79.08	83.3	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	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	59.2	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

NOTE: Table generated using SAS, a statistical software package. Percentiles of the Chi-square distribution $\chi_{p,\upsilon}$ are listed for various degrees of freedom υ: $p = P(\chi_{\upsilon} \leq \chi_{p,\upsilon})$.

Table B-3.
Values of the Parameter λ for Cohen's Estimates

							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

Source: EPA/600/R-96/084.

Table B-4.
Critical Values of *D* for the Discordance Test

n	**Level of Significance, α**	
	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
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

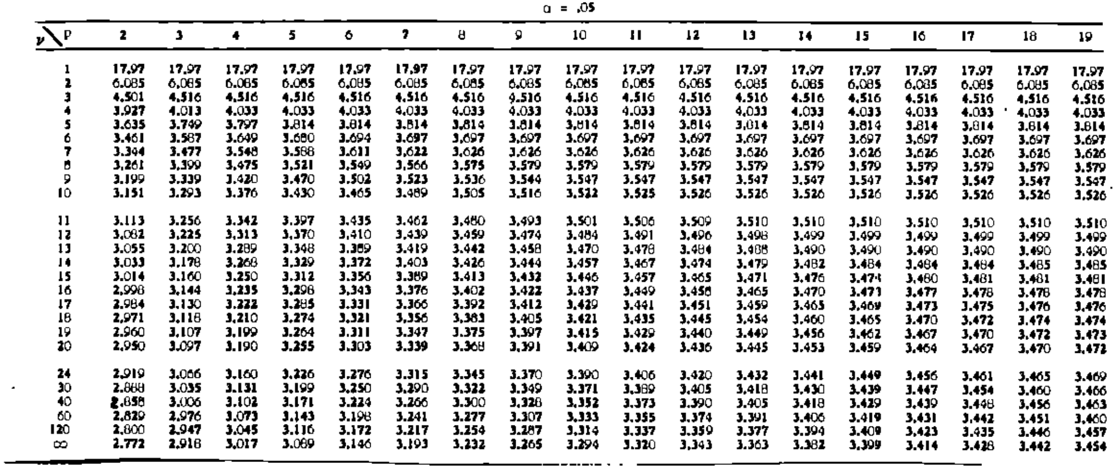
Source: EPA/600/R-96/084.

Table B-5.
Critical Values for Dixon's Test (Extreme Value Test)

n	**Level of Significance, α**		
	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

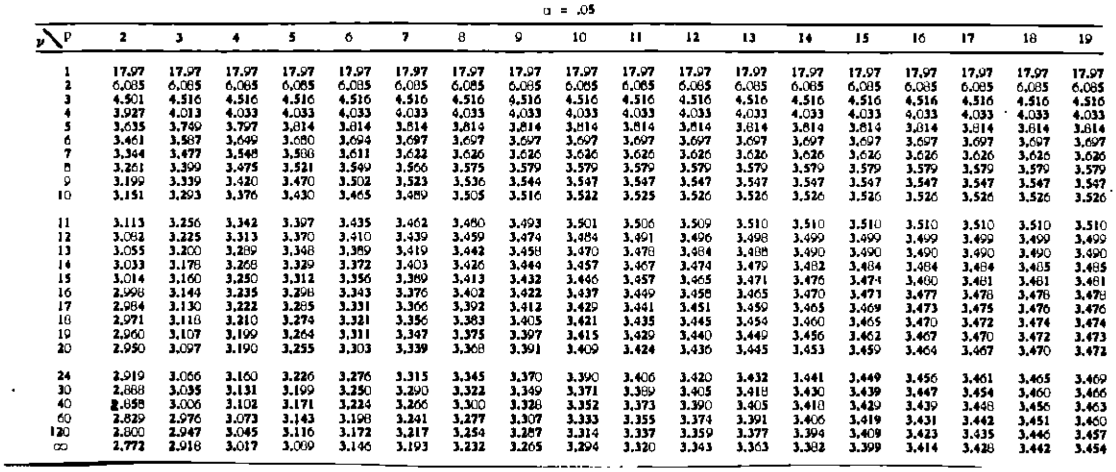
Source: EPA/600/R-96/084.

Table B-6.
Critical Values for Duncan's Multiple Range Test



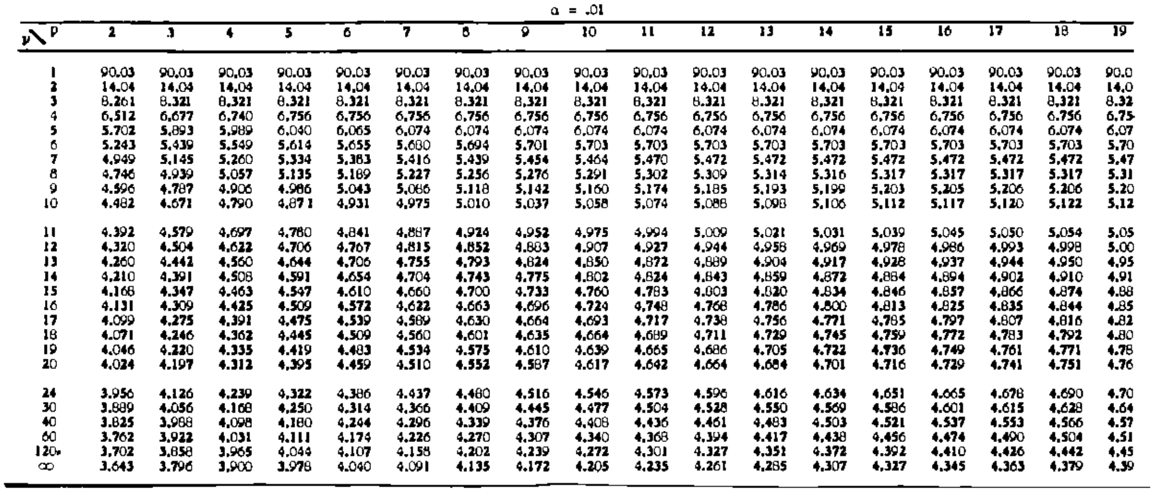
α = .05

ν\P	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97
2	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085
3	4.501	4.516	4.516	4.516	4.516	4.516	4.516	9.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	3.927	4.013	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.635	3.749	3.797	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814
6	3.461	3.587	3.649	3.680	3.694	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697
7	3.344	3.477	3.548	3.588	3.611	3.622	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626
8	3.261	3.399	3.475	3.521	3.549	3.566	3.575	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522	3.525	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526
11	3.113	3.256	3.342	3.397	3.435	3.462	3.480	3.493	3.501	3.506	3.509	3.510	3.510	3.510	3.510	3.510	3.510	3.510
12	3.082	3.225	3.313	3.370	3.410	3.439	3.459	3.474	3.484	3.491	3.496	3.498	3.499	3.499	3.499	3.499	3.499	3.499
13	3.055	3.200	3.289	3.348	3.389	3.419	3.442	3.458	3.470	3.478	3.484	3.488	3.490	3.490	3.490	3.490	3.490	3.490
14	3.033	3.178	3.268	3.329	3.372	3.403	3.426	3.444	3.457	3.467	3.474	3.479	3.482	3.484	3.484	3.484	3.485	3.485
15	3.014	3.160	3.250	3.312	3.356	3.389	3.413	3.432	3.446	3.457	3.465	3.471	3.476	3.474	3.480	3.481	3.481	3.481
16	2.998	3.144	3.235	3.298	3.343	3.376	3.402	3.422	3.437	3.449	3.458	3.465	3.470	3.473	3.477	3.478	3.478	3.478
17	2.984	3.130	3.222	3.285	3.331	3.366	3.392	3.412	3.429	3.441	3.451	3.459	3.465	3.469	3.473	3.475	3.476	3.476
18	2.971	3.118	3.210	3.274	3.321	3.356	3.383	3.405	3.421	3.435	3.445	3.454	3.460	3.465	3.470	3.472	3.474	3.474
19	2.960	3.107	3.199	3.264	3.311	3.347	3.375	3.397	3.415	3.429	3.440	3.449	3.456	3.462	3.467	3.470	3.472	3.473
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.391	3.409	3.424	3.436	3.445	3.453	3.459	3.464	3.467	3.470	3.472
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390	3.406	3.420	3.432	3.441	3.449	3.456	3.461	3.465	3.469
30	2.888	3.035	3.131	3.199	3.250	3.290	3.322	3.349	3.371	3.389	3.405	3.418	3.430	3.439	3.447	3.454	3.460	3.466
40	2.858	3.006	3.102	3.171	3.224	3.266	3.300	3.328	3.352	3.373	3.390	3.405	3.418	3.429	3.439	3.448	3.456	3.463
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333	3.355	3.374	3.391	3.406	3.419	3.431	3.442	3.451	3.460
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.287	3.314	3.337	3.359	3.377	3.394	3.409	3.423	3.435	3.446	3.457
∞	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294	3.320	3.343	3.363	3.382	3.399	3.414	3.428	3.442	3.454



α = .05

ν\P	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97
2	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085
3	4.501	4.516	4.516	4.516	4.516	4.516	4.516	9.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	3.927	4.013	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.635	3.749	3.797	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814
6	3.461	3.587	3.649	3.680	3.694	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697
7	3.344	3.477	3.548	3.588	3.611	3.622	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626
8	3.261	3.399	3.475	3.521	3.549	3.566	3.575	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522	3.525	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526
11	3.113	3.256	3.342	3.397	3.435	3.462	3.480	3.493	3.501	3.506	3.509	3.510	3.510	3.510	3.510	3.510	3.510	3.510
12	3.082	3.225	3.313	3.370	3.410	3.439	3.459	3.474	3.484	3.491	3.496	3.498	3.499	3.499	3.499	3.499	3.499	3.499
13	3.055	3.200	3.289	3.348	3.389	3.419	3.442	3.458	3.470	3.478	3.484	3.488	3.490	3.490	3.490	3.490	3.490	3.490
14	3.033	3.178	3.268	3.329	3.372	3.403	3.426	3.444	3.457	3.467	3.474	3.479	3.482	3.484	3.484	3.484	3.485	3.485
15	3.014	3.160	3.250	3.312	3.356	3.389	3.413	3.432	3.446	3.457	3.465	3.471	3.476	3.474	3.480	3.481	3.481	3.481
16	2.998	3.144	3.235	3.298	3.343	3.376	3.402	3.422	3.437	3.449	3.458	3.465	3.470	3.473	3.477	3.478	3.478	3.478
17	2.984	3.130	3.222	3.285	3.331	3.366	3.392	3.412	3.429	3.441	3.451	3.459	3.465	3.469	3.473	3.475	3.476	3.476
18	2.971	3.118	3.210	3.274	3.321	3.356	3.383	3.405	3.421	3.435	3.445	3.454	3.460	3.465	3.470	3.472	3.474	3.474
19	2.960	3.107	3.199	3.264	3.311	3.347	3.375	3.397	3.415	3.429	3.440	3.449	3.456	3.462	3.467	3.470	3.472	3.473
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.391	3.409	3.424	3.436	3.445	3.453	3.459	3.464	3.467	3.470	3.472
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390	3.406	3.420	3.432	3.441	3.449	3.456	3.461	3.465	3.469
30	2.888	3.035	3.131	3.199	3.250	3.290	3.322	3.349	3.371	3.389	3.405	3.418	3.430	3.439	3.447	3.454	3.460	3.466
40	2.858	3.006	3.102	3.171	3.224	3.266	3.300	3.328	3.352	3.373	3.390	3.405	3.418	3.429	3.439	3.448	3.456	3.463
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333	3.355	3.374	3.391	3.406	3.419	3.431	3.442	3.451	3.460
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.287	3.314	3.337	3.359	3.377	3.394	3.409	3.423	3.435	3.446	3.457
∞	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294	3.320	3.343	3.363	3.382	3.399	3.414	3.428	3.442	3.454



α = .01

ν\P	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.0
2	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.0
3	8.261	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.32
4	6.512	6.677	6.740	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.75
5	5.702	5.893	5.989	6.040	6.065	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.07
6	5.243	5.439	5.549	5.614	5.655	5.680	5.694	5.701	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.70
7	4.949	5.145	5.260	5.334	5.383	5.416	5.439	5.454	5.464	5.470	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.47
8	4.746	4.939	5.057	5.135	5.189	5.227	5.256	5.276	5.291	5.302	5.309	5.314	5.316	5.317	5.317	5.317	5.317	5.31
9	4.596	4.787	4.906	4.986	5.043	5.086	5.118	5.142	5.160	5.174	5.185	5.193	5.199	5.203	5.205	5.206	5.206	5.20
10	4.482	4.671	4.790	4.871	4.931	4.975	5.010	5.037	5.058	5.074	5.088	5.098	5.106	5.112	5.117	5.120	5.122	5.12
11	4.392	4.579	4.697	4.780	4.841	4.887	4.924	4.952	4.975	4.994	5.009	5.021	5.031	5.039	5.045	5.050	5.054	5.05
12	4.320	4.504	4.622	4.706	4.767	4.815	4.852	4.883	4.907	4.927	4.944	4.958	4.969	4.978	4.986	4.993	4.998	5.00
13	4.260	4.442	4.560	4.644	4.706	4.755	4.793	4.824	4.850	4.872	4.889	4.904	4.917	4.928	4.937	4.944	4.950	4.95
14	4.210	4.391	4.508	4.591	4.654	4.704	4.743	4.775	4.802	4.824	4.843	4.859	4.872	4.884	4.894	4.902	4.910	4.91
15	4.168	4.347	4.463	4.547	4.610	4.660	4.700	4.733	4.760	4.783	4.803	4.820	4.834	4.846	4.857	4.866	4.874	4.88
16	4.131	4.309	4.425	4.509	4.572	4.622	4.663	4.696	4.724	4.748	4.768	4.786	4.800	4.813	4.825	4.835	4.844	4.85
17	4.099	4.275	4.391	4.475	4.539	4.589	4.630	4.664	4.693	4.717	4.738	4.756	4.771	4.785	4.797	4.807	4.816	4.82
18	4.071	4.246	4.362	4.445	4.509	4.560	4.601	4.635	4.664	4.689	4.711	4.729	4.745	4.759	4.772	4.783	4.792	4.80
19	4.046	4.220	4.335	4.419	4.483	4.534	4.575	4.610	4.639	4.665	4.686	4.705	4.722	4.736	4.749	4.761	4.771	4.78
20	4.024	4.197	4.312	4.395	4.459	4.510	4.552	4.587	4.617	4.642	4.664	4.684	4.701	4.716	4.729	4.741	4.751	4.76
24	3.956	4.126	4.239	4.322	4.386	4.437	4.480	4.516	4.546	4.573	4.596	4.616	4.634	4.651	4.665	4.678	4.690	4.70
30	3.889	4.056	4.168	4.250	4.314	4.366	4.409	4.445	4.477	4.504	4.528	4.550	4.569	4.586	4.601	4.615	4.628	4.64
40	3.825	3.988	4.098	4.180	4.244	4.296	4.339	4.376	4.408	4.436	4.461	4.483	4.503	4.521	4.537	4.553	4.566	4.57
60	3.762	3.922	4.031	4.111	4.174	4.226	4.270	4.307	4.340	4.368	4.394	4.417	4.438	4.456	4.474	4.490	4.504	4.51
120	3.702	3.858	3.965	4.044	4.107	4.158	4.202	4.239	4.272	4.301	4.327	4.351	4.372	4.392	4.410	4.426	4.442	4.45
∞	3.643	3.796	3.900	3.978	4.040	4.091	4.135	4.172	4.205	4.235	4.261	4.285	4.307	4.327	4.345	4.363	4.379	4.39

α = .01

ν\p	20	22	24	26	28	30	32	34	36	38	40	50	60	70	80	90	100
1	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03
2	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04
3	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321
4	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756
5	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074
6	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703
7	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472
8	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317
9	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.026	5.206	5.206
10	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124
11	5.059	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061
12	5.006	5.010	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011
13	4.960	4.966	4.970	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972
14	4.921	4.929	4.935	4.938	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940
15	4.887	4.897	4.904	4.909	4.912	4.914	4.914	4.914	4.914	4.914	4.914	4.914	4.914	4.914	4.914	4.914	4.914
16	4.858	4.869	4.877	4.883	4.887	4.890	4.892	4.892	4.892	4.892	4.892	4.892	4.892	4.892	4.892	4.892	4.892
17	4.832	4.844	4.853	4.860	4.865	4.869	4.872	4.873	4.874	4.874	4.874	4.874	4.874	4.874	4.874	4.874	4.874
18	4.808	4.821	4.832	4.839	4.846	4.850	4.854	4.856	4.857	4.858	4.858	4.858	4.858	4.858	4.858	4.858	4.858
19	4.788	4.802	4.812	4.821	4.828	4.833	4.838	4.841	4.843	4.844	4.845	4.845	4.845	4.845	4.845	4.845	4.845
20	4.769	4.786	4.795	4.805	4.813	4.818	4.823	4.827	4.830	4.832	4.833	4.833	4.833	4.833	4.833	4.833	4.833
24	4.710	4.727	4.741	4.752	4.762	4.770	4.777	4.783	4.788	4.791	4.794	4.802	4.802	4.802	4.802	4.802	4.802
30	4.650	4.669	4.685	4.699	4.711	4.721	4.730	4.738	4.744	4.750	4.755	4.772	4.777	4.777	4.777	4.777	4.777
40	4.591	4.611	4.630	4.645	4.659	4.671	4.682	4.692	4.700	4.708	4.715	4.740	4.754	4.761	4.764	4.764	4.764
60	4.530	4.553	4.573	4.591	4.607	4.620	4.633	4.645	4.655	4.665	4.673	4.707	4.730	4.745	4.755	4.761	4.665
120	4.469	4.494	4.516	4.535	4.552	4.568	4.583	4.596	4.609	4.619	4.630	4.673	4.703	4.727	4.745	4.759	4.770
∞	4.408	4.434	4.457	4.478	4.497	4.514	4.530	4.545	4.559	4.572	4.584	4.635	4.675	4.707	4.734	4.756	4.776

Source: Reproduced from H. L. Harter, "Critical Values for Duncan's Multiple Range Test." *Biometrics*, **16**, 671–685 (1960). With permission from the Biometric Society.

Source: Mason et al. (1989).

Table B-7.
Percentiles of the F Distribution

α = .01

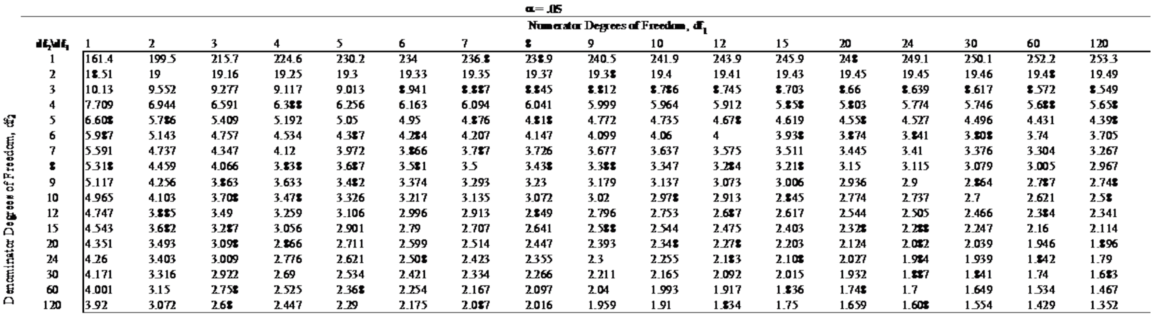
Denominator Degrees of Freedom, df_2

	Numerator Degrees of Freedom, df_1																
df_2\\df_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	60	120
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6313	6339
2	98.5	99	99.17	99.25	99.3	99.33	99.36	99.37	99.39	99.4	99.42	99.43	99.45	99.46	99.47	99.48	99.49
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.6	26.5	26.32	26.22
4	21.2	18	16.69	15.98	15.52	15.21	14.98	14.8	14.66	14.55	14.37	14.2	14.02	13.93	13.84	13.65	13.56
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.888	9.722	9.553	9.466	9.379	9.202	9.112
6	13.75	10.92	9.78	9.148	8.746	8.466	8.26	8.102	7.976	7.874	7.718	7.559	7.396	7.313	7.229	7.057	6.969
7	12.25	9.547	8.451	7.847	7.46	7.191	6.993	6.84	6.719	6.62	6.469	6.314	6.155	6.074	5.992	5.824	5.737
8	11.26	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.515	5.359	5.279	5.198	5.032	4.946
9	10.56	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.962	4.808	4.729	4.649	4.483	4.398
10	10.04	7.559	6.552	5.994	5.636	5.386	5.2	5.057	4.942	4.849	4.706	4.558	4.405	4.327	4.247	4.082	3.996
12	9.33	6.927	5.953	5.412	5.064	4.821	4.64	4.499	4.388	4.296	4.155	4.01	3.858	3.78	3.701	3.535	3.449
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.522	3.372	3.294	3.214	3.047	2.959
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	3.088	2.938	2.859	2.778	2.608	2.517
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.889	2.738	2.659	2.577	2.403	2.31
30	7.562	5.39	4.51	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.843	2.7	2.549	2.469	2.386	2.208	2.111
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.496	2.352	2.198	2.115	2.028	1.836	1.726
120	6.851	4.787	3.949	3.48	3.174	2.956	2.792	2.663	2.559	2.472	2.336	2.192	2.035	1.95	1.86	1.656	1.533

α = .025

Denominator Degrees of Freedom, df_2

	Numerator Degrees of Freedom, df_1																
df_2\\df_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	60	120
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1010	1014
2	38.51	39	39.17	39.25	39.3	39.33	39.36	39.37	39.39	39.4	39.41	39.43	39.45	39.46	39.46	39.48	39.49
3	17.44	16.04	15.44	15.1	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	13.99	13.95
4	12.22	10.65	9.979	9.605	9.364	9.197	9.074	8.98	8.905	8.844	8.751	8.657	8.56	8.511	8.461	8.36	8.309
5	10.01	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681	6.619	6.525	6.428	6.329	6.278	6.227	6.123	6.069
6	8.813	7.26	6.599	6.227	5.988	5.82	5.695	5.6	5.523	5.461	5.366	5.269	5.168	5.117	5.065	4.959	4.904
7	8.073	6.542	5.89	5.523	5.285	5.119	4.995	4.899	4.823	4.761	4.666	4.568	4.467	4.415	4.362	4.254	4.199
8	7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433	4.357	4.295	4.2	4.101	3.999	3.947	3.894	3.784	3.728
9	7.209	5.715	5.078	4.718	4.484	4.32	4.197	4.102	4.026	3.964	3.868	3.769	3.667	3.614	3.56	3.449	3.392
10	6.937	5.456	4.826	4.468	4.236	4.072	3.95	3.855	3.779	3.717	3.621	3.522	3.419	3.365	3.311	3.198	3.14
12	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436	3.374	3.277	3.177	3.073	3.019	2.963	2.848	2.787
15	6.2	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123	3.06	2.963	2.862	2.756	2.701	2.644	2.524	2.461
20	5.871	4.461	3.859	3.515	3.289	3.128	3.007	2.913	2.837	2.774	2.676	2.573	2.464	2.408	2.349	2.223	2.156
24	5.717	4.319	3.721	3.379	3.155	2.995	2.874	2.779	2.703	2.64	2.541	2.437	2.327	2.269	2.209	2.08	2.01
30	5.568	4.182	3.589	3.25	3.026	2.867	2.746	2.651	2.575	2.511	2.412	2.307	2.195	2.136	2.074	1.94	1.866
60	5.286	3.925	3.343	3.008	2.786	2.627	2.507	2.412	2.334	2.27	2.169	2.061	1.944	1.882	1.815	1.667	1.581
120	5.152	3.805	3.227	2.894	2.674	2.515	2.395	2.299	2.222	2.157	2.055	1.945	1.825	1.76	1.69	1.53	1.433

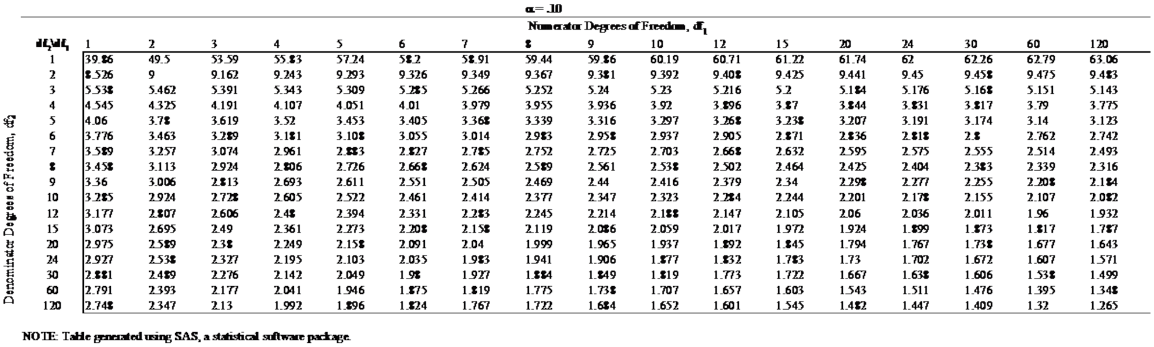


α= .05

Numerator Degrees of Freedom, df_1

Denominator Degrees of Freedom, df_2

df_2\df_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	60	120
1	161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	240.5	241.9	243.9	245.9	248	249.1	250.1	252.2	253.3
2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	19.41	19.43	19.45	19.45	19.46	19.48	19.49
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.745	8.703	8.66	8.639	8.617	8.572	8.549
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.858	5.803	5.774	5.746	5.688	5.658
5	6.608	5.786	5.409	5.192	5.05	4.95	4.876	4.818	4.772	4.735	4.678	4.619	4.558	4.527	4.496	4.431	4.398
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.06	4	3.938	3.874	3.841	3.808	3.74	3.705
7	5.591	4.737	4.347	4.12	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.511	3.445	3.41	3.376	3.304	3.267
8	5.318	4.459	4.066	3.838	3.687	3.581	3.5	3.438	3.388	3.347	3.284	3.218	3.15	3.115	3.079	3.005	2.967
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.23	3.179	3.137	3.073	3.006	2.936	2.9	2.864	2.787	2.748
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.02	2.978	2.913	2.845	2.774	2.737	2.7	2.621	2.58
12	4.747	3.885	3.49	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.617	2.544	2.505	2.466	2.384	2.341
15	4.543	3.682	3.287	3.056	2.901	2.79	2.707	2.641	2.588	2.544	2.475	2.403	2.328	2.288	2.247	2.16	2.114
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.203	2.124	2.082	2.039	1.946	1.896
24	4.26	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.3	2.255	2.183	2.108	2.027	1.984	1.939	1.842	1.79
30	4.171	3.316	2.922	2.69	2.534	2.421	2.334	2.266	2.211	2.165	2.092	2.015	1.932	1.887	1.841	1.74	1.683
60	4.001	3.15	2.758	2.525	2.368	2.254	2.167	2.097	2.04	1.993	1.917	1.836	1.748	1.7	1.649	1.534	1.467
120	3.92	3.072	2.68	2.447	2.29	2.175	2.087	2.016	1.959	1.91	1.834	1.75	1.659	1.608	1.554	1.429	1.352



α= .10

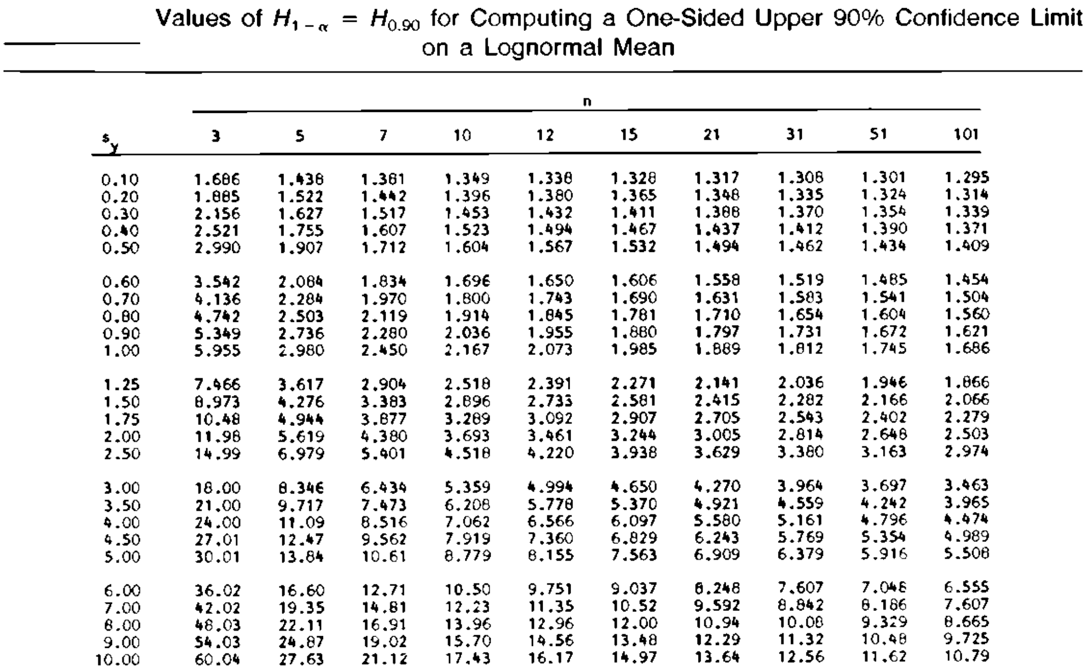
Numerator Degrees of Freedom, df_1

Denominator Degrees of Freedom, df_2

df_2\df_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	60	120
1	39.86	49.5	53.59	55.83	57.24	58.2	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62	62.26	62.79	63.06
2	8.526	9	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392	9.408	9.425	9.441	9.45	9.458	9.475	9.483
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.24	5.23	5.216	5.2	5.184	5.176	5.168	5.151	5.143
4	4.545	4.325	4.191	4.107	4.051	4.01	3.979	3.955	3.936	3.92	3.896	3.87	3.844	3.831	3.817	3.79	3.775
5	4.06	3.78	3.619	3.52	3.453	3.405	3.368	3.339	3.316	3.297	3.268	3.238	3.207	3.191	3.174	3.14	3.123
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937	2.905	2.871	2.836	2.818	2.8	2.762	2.742
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703	2.668	2.632	2.595	2.575	2.555	2.514	2.493
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538	2.502	2.464	2.425	2.404	2.383	2.339	2.316
9	3.36	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.44	2.416	2.379	2.34	2.298	2.277	2.255	2.208	2.184
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323	2.284	2.244	2.201	2.178	2.155	2.107	2.082
12	3.177	2.807	2.606	2.48	2.394	2.331	2.283	2.245	2.214	2.188	2.147	2.105	2.06	2.036	2.011	1.96	1.932
15	3.073	2.695	2.49	2.361	2.273	2.208	2.158	2.119	2.086	2.059	2.017	1.972	1.924	1.899	1.873	1.817	1.787
20	2.975	2.589	2.38	2.249	2.158	2.091	2.04	1.999	1.965	1.937	1.892	1.845	1.794	1.767	1.738	1.677	1.643
24	2.927	2.538	2.327	2.195	2.103	2.035	1.983	1.941	1.906	1.877	1.832	1.783	1.73	1.702	1.672	1.607	1.571
30	2.881	2.489	2.276	2.142	2.049	1.98	1.927	1.884	1.849	1.819	1.773	1.722	1.667	1.638	1.606	1.538	1.499
60	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707	1.657	1.603	1.543	1.511	1.476	1.395	1.348
120	2.748	2.347	2.13	1.992	1.896	1.824	1.767	1.722	1.684	1.652	1.601	1.545	1.482	1.447	1.409	1.32	1.265

NOTE: Table generated using SAS, a statistical software package.

Table B-8.
H-statistic for Confidence Limit on a Lognormal Mean



Values of $H_{1-\alpha} = H_{0.90}$ for Computing a One-Sided Upper 90% Confidence Limit on a Lognormal Mean

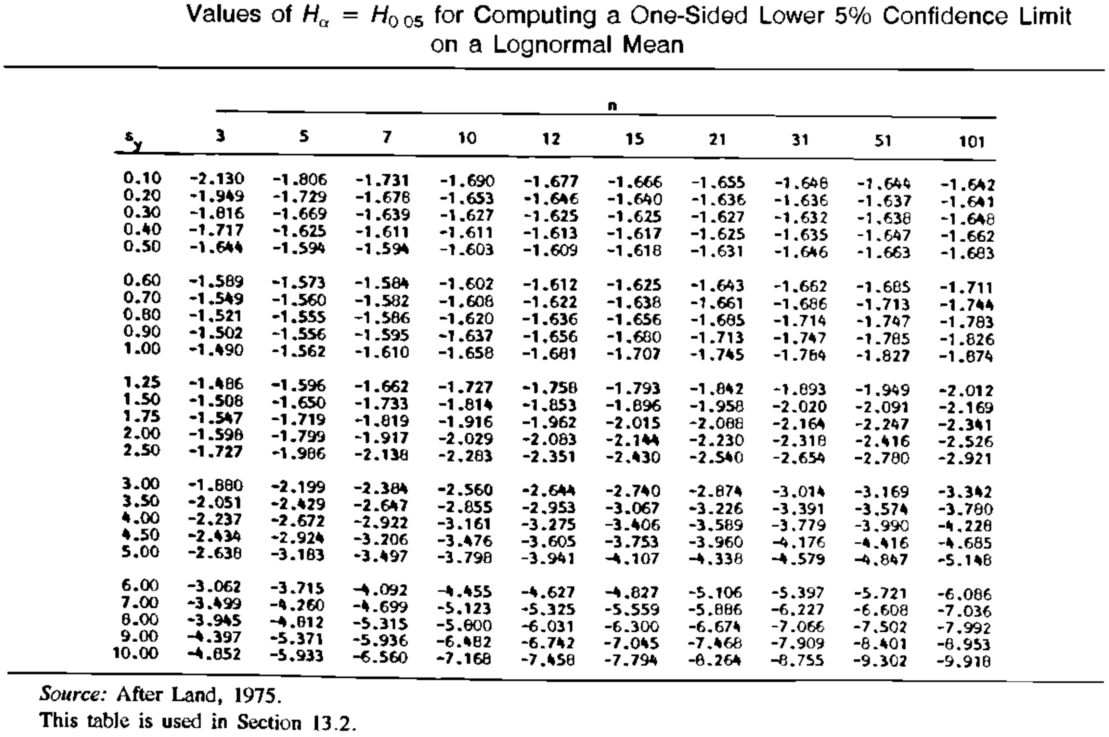
	n									
s_y	3	5	7	10	12	15	21	31	51	101
0.10	1.686	1.438	1.381	1.349	1.338	1.328	1.317	1.308	1.301	1.295
0.20	1.885	1.522	1.442	1.396	1.380	1.365	1.348	1.335	1.324	1.314
0.30	2.156	1.627	1.517	1.453	1.432	1.411	1.388	1.370	1.354	1.339
0.40	2.521	1.755	1.607	1.523	1.494	1.467	1.437	1.412	1.390	1.371
0.50	2.990	1.907	1.712	1.604	1.567	1.532	1.494	1.462	1.434	1.409
0.60	3.542	2.084	1.834	1.696	1.650	1.606	1.558	1.519	1.485	1.454
0.70	4.136	2.284	1.970	1.800	1.743	1.690	1.631	1.583	1.541	1.504
0.80	4.742	2.503	2.119	1.914	1.845	1.781	1.710	1.654	1.604	1.560
0.90	5.349	2.736	2.280	2.036	1.955	1.880	1.797	1.731	1.672	1.621
1.00	5.955	2.980	2.450	2.167	2.073	1.985	1.889	1.812	1.745	1.686
1.25	7.466	3.617	2.904	2.518	2.391	2.271	2.141	2.036	1.946	1.866
1.50	8.973	4.276	3.383	2.896	2.733	2.581	2.415	2.282	2.166	2.066
1.75	10.48	4.944	3.877	3.289	3.092	2.907	2.705	2.543	2.402	2.279
2.00	11.98	5.619	4.380	3.693	3.461	3.244	3.005	2.814	2.648	2.503
2.50	14.99	6.979	5.401	4.518	4.220	3.938	3.629	3.380	3.163	2.974
3.00	18.00	8.346	6.434	5.359	4.994	4.650	4.270	3.964	3.697	3.463
3.50	21.00	9.717	7.473	6.208	5.778	5.370	4.921	4.559	4.242	3.965
4.00	24.00	11.09	8.516	7.062	6.566	6.097	5.580	5.161	4.796	4.474
4.50	27.01	12.47	9.562	7.919	7.360	6.829	6.243	5.769	5.354	4.989
5.00	30.01	13.84	10.61	8.779	8.155	7.563	6.909	6.379	5.916	5.508
6.00	36.02	16.60	12.71	10.50	9.751	9.037	8.248	7.607	7.048	6.555
7.00	42.02	19.35	14.81	12.23	11.35	10.52	9.592	8.842	8.186	7.607
8.00	48.03	22.11	16.91	13.96	12.96	12.00	10.94	10.08	9.329	8.665
9.00	54.03	24.87	19.02	15.70	14.56	13.48	12.29	11.32	10.48	9.725
10.00	60.04	27.63	21.12	17.43	16.17	14.97	13.64	12.56	11.62	10.79

Values of $H_\alpha = H_{0.10}$ for Computing a One-Sided Lower 10% Confidence Limit on a Lognormal Mean

	n									
s_y	3	5	7	10	12	15	21	31	51	101
0.10	-1.431	-1.320	-1.296	-1.285	-1.281	-1.279	-1.277	-1.277	-1.278	-1.279
0.20	-1.350	-1.281	-1.268	-1.266	-1.266	-1.266	-1.268	-1.272	-1.275	-1.280
0.30	-1.289	-1.252	-1.250	-1.254	-1.257	-1.260	-1.266	-1.272	-1.280	-1.287
0.40	-1.245	-1.233	-1.239	-1.249	-1.254	-1.261	-1.270	-1.279	-1.289	-1.301
0.50	-1.213	-1.221	-1.234	-1.250	-1.257	-1.266	-1.279	-1.291	-1.304	-1.319
0.60	-1.190	-1.215	-1.235	-1.256	-1.266	-1.277	-1.292	-1.307	-1.324	-1.342
0.70	-1.176	-1.215	-1.241	-1.266	-1.278	-1.292	-1.310	-1.329	-1.349	-1.370
0.80	-1.168	-1.219	-1.251	-1.280	-1.294	-1.311	-1.332	-1.354	-1.377	-1.403
0.90	-1.165	-1.227	-1.264	-1.298	-1.314	-1.333	-1.358	-1.383	-1.409	-1.439
1.00	-1.166	-1.239	-1.281	-1.320	-1.337	-1.358	-1.387	-1.414	-1.445	-1.478
1.25	-1.184	-1.280	-1.334	-1.384	-1.407	-1.434	-1.470	-1.507	-1.547	-1.589
1.50	-1.217	-1.334	-1.400	-1.462	-1.491	-1.523	-1.568	-1.613	-1.063	-1.716
1.75	-1.260	-1.398	-1.477	-1.551	-1.585	-1.624	-1.677	-1.732	-1.790	-1.855
2.00	-1.310	-1.470	-1.562	-1.647	-1.688	-1.733	-1.795	-1.859	-1.928	-2.003
2.50	-1.426	-1.634	-1.751	-1.862	-1.913	-1.971	-2.051	-2.133	-2.223	-2.321
3.00	-1.560	-1.817	-1.960	-2.095	-2.157	-2.229	-2.326	-2.427	-2.536	-2.657
3.50	-1.710	-2.014	-2.183	-2.341	-2.415	-2.499	-2.615	-2.733	-2.864	-3.007
4.00	-1.871	-2.221	-2.415	-2.596	-2.681	-2.778	-2.913	-3.050	-3.200	-3.366
4.50	-2.041	-2.435	-2.653	-2.858	-2.955	-3.064	-3.217	-3.372	-3.542	-3.731
5.00	-2.217	-2.654	-2.897	-3.126	-3.233	-3.356	-3.525	-3.698	-3.889	-4.100
6.00	-2.581	-3.104	-3.396	-3.671	-3.800	-3.949	-4.153	-4.363	-4.594	-4.849
7.00	-2.955	-3.564	-3.904	-4.226	-4.377	-4.549	-4.790	-5.037	-5.307	-5.607
8.00	-3.336	-4.030	-4.418	-4.787	-4.960	-5.159	-5.433	-5.715	-6.026	-6.370
9.00	-3.721	-4.500	-4.937	-5.352	-5.547	-5.771	-6.080	-6.399	-6.748	-7.136
10.00	-4.109	-4.973	-5.459	-5.920	-6.137	-6.386	-6.730	-7.085	-7.474	-7.906

Values of $H_{1-\alpha} = H_{0.95}$ for Computing a One-Sided Upper 95% Confidence Limit on a Lognormal Mean

	n									
s_y	3	5	7	10	12	15	21	31	51	101
0.10	2.750	2.035	1.886	1.802	1.775	1.749	1.722	1.701	1.684	1.670
0.20	3.295	2.198	1.992	1.881	1.843	1.809	1.771	1.742	1.718	1.697
0.30	4.109	2.402	2.125	1.977	1.927	1.882	1.833	1.793	1.761	1.733
0.40	5.220	2.651	2.282	2.089	2.026	1.968	1.905	1.856	1.813	1.777
0.50	6.495	2.947	2.465	2.220	2.141	2.068	1.989	1.928	1.876	1.830
0.60	7.807	3.287	2.673	2.368	2.271	2.181	2.085	2.010	1.946	1.891
0.70	9.120	3.662	2.904	2.532	2.414	2.306	2.191	2.102	2.025	1.960
0.80	10.43	4.062	3.155	2.710	2.570	2.443	2.307	2.202	2.112	2.035
0.90	11.74	4.478	3.420	2.902	2.738	2.589	2.432	2.310	2.206	2.117
1.00	13.05	4.905	3.696	3.103	2.915	2.744	2.564	2.423	2.306	2.205
1.25	16.33	6.001	4.426	3.639	3.389	3.163	2.923	2.737	2.580	2.447
1.50	19.60	7.120	5.184	4.207	3.896	3.612	3.311	3.077	2.881	2.713
1.75	22.87	8.250	5.960	4.795	4.422	4.081	3.719	3.437	3.200	2.997
2.00	26.14	9.387	6.747	5.396	4.962	4.564	4.141	3.812	3.533	3.295
2.50	32.69	11.67	8.339	6.621	6.067	5.557	5.013	4.588	4.228	3.920
3.00	39.23	13.97	9.945	7.864	7.191	6.570	5.907	5.388	4.947	4.569
3.50	45.77	16.27	11.56	9.118	8.326	7.596	6.815	6.201	5.681	5.233
4.00	52.31	18.58	13.18	10.38	9.469	8.630	7.731	7.024	6.424	5.908
4.50	58.85	20.88	14.80	11.64	10.62	9.669	8.652	7.854	7.174	6.590
5.00	65.39	23.19	16.43	12.91	11.77	10.71	9.579	8.688	7.929	7.277
6.00	78.47	27.81	19.68	15.45	14.08	12.81	11.44	10.36	9.449	8.661
7.00	91.55	32.43	22.94	18.00	16.39	14.90	13.31	12.05	10.98	10.05
8.00	104.6	37.06	26.20	20.55	18.71	17.01	15.18	13.74	12.51	11.45
9.00	117.7	41.68	29.46	23.10	21.03	19.11	17.05	15.43	14.05	12.85
10.00	130.8	46.31	32.73	25.66	23.35	21.22	18.93	17.13	15.59	14.26



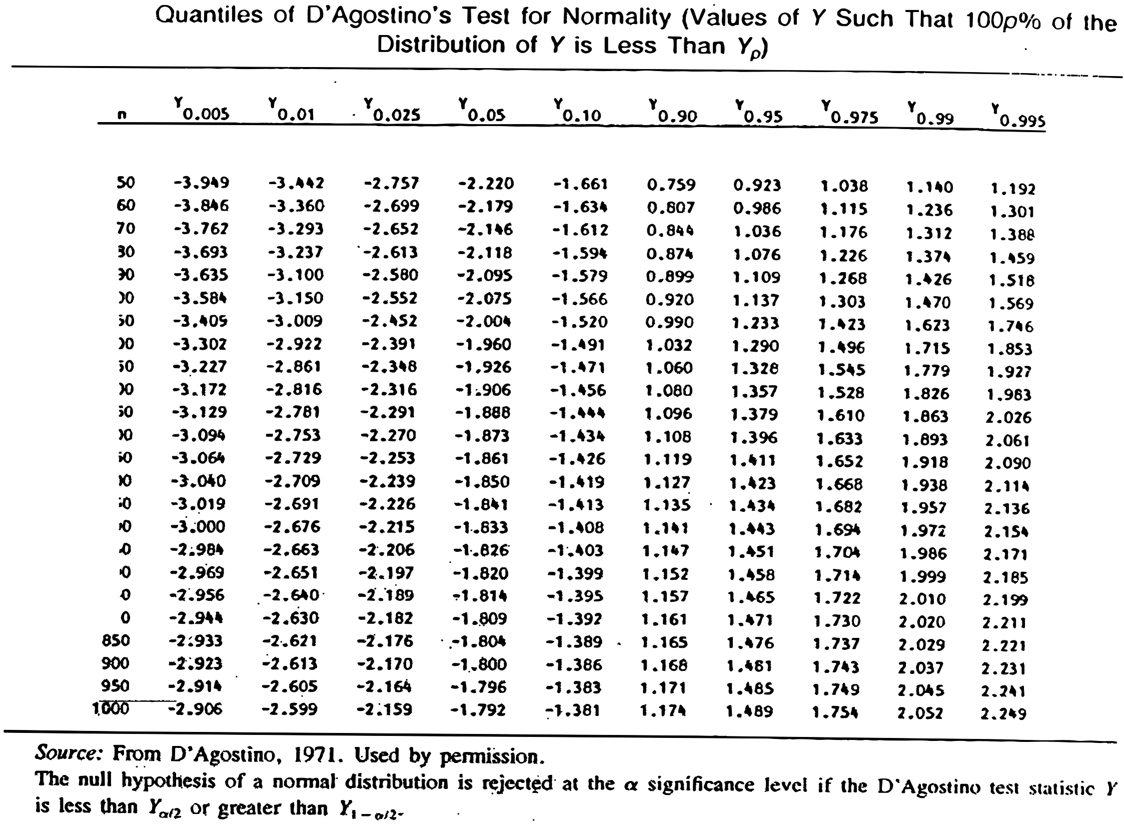
Values of $H_{\alpha} = H_{0\,05}$ for Computing a One-Sided Lower 5% Confidence Limit on a Lognormal Mean

	n									
s_y	3	5	7	10	12	15	21	31	51	101
0.10	-2.130	-1.806	-1.731	-1.690	-1.677	-1.666	-1.655	-1.648	-1.644	-1.642
0.20	-1.949	-1.729	-1.678	-1.653	-1.646	-1.640	-1.636	-1.636	-1.637	-1.641
0.30	-1.816	-1.669	-1.639	-1.627	-1.625	-1.625	-1.627	-1.632	-1.638	-1.648
0.40	-1.717	-1.625	-1.611	-1.611	-1.613	-1.617	-1.625	-1.635	-1.647	-1.662
0.50	-1.644	-1.594	-1.594	-1.603	-1.609	-1.618	-1.631	-1.646	-1.663	-1.683
0.60	-1.589	-1.573	-1.584	-1.602	-1.612	-1.625	-1.643	-1.662	-1.685	-1.711
0.70	-1.549	-1.560	-1.582	-1.608	-1.622	-1.638	-1.661	-1.686	-1.713	-1.744
0.80	-1.521	-1.555	-1.586	-1.620	-1.636	-1.656	-1.685	-1.714	-1.747	-1.783
0.90	-1.502	-1.556	-1.595	-1.637	-1.656	-1.680	-1.713	-1.747	-1.785	-1.826
1.00	-1.490	-1.562	-1.610	-1.658	-1.681	-1.707	-1.745	-1.764	-1.827	-1.874
1.25	-1.486	-1.596	-1.662	-1.727	-1.758	-1.793	-1.842	-1.893	-1.949	-2.012
1.50	-1.508	-1.650	-1.733	-1.814	-1.853	-1.896	-1.958	-2.020	-2.091	-2.169
1.75	-1.547	-1.719	-1.819	-1.916	-1.962	-2.015	-2.088	-2.164	-2.247	-2.341
2.00	-1.598	-1.799	-1.917	-2.029	-2.083	-2.144	-2.230	-2.318	-2.416	-2.526
2.50	-1.727	-1.986	-2.138	-2.283	-2.351	-2.430	-2.540	-2.654	-2.780	-2.921
3.00	-1.880	-2.199	-2.384	-2.560	-2.644	-2.740	-2.874	-3.014	-3.169	-3.342
3.50	-2.051	-2.429	-2.647	-2.855	-2.953	-3.067	-3.226	-3.391	-3.574	-3.780
4.00	-2.237	-2.672	-2.922	-3.161	-3.275	-3.406	-3.589	-3.779	-3.990	-4.228
4.50	-2.434	-2.924	-3.206	-3.476	-3.605	-3.753	-3.960	-4.176	-4.416	-4.685
5.00	-2.638	-3.183	-3.497	-3.798	-3.941	-4.107	-4.338	-4.579	-4.847	-5.148
6.00	-3.062	-3.715	-4.092	-4.455	-4.627	-4.827	-5.106	-5.397	-5.721	-6.086
7.00	-3.499	-4.260	-4.699	-5.123	-5.325	-5.559	-5.886	-6.227	-6.608	-7.036
8.00	-3.945	-4.812	-5.315	-5.800	-6.031	-6.300	-6.674	-7.066	-7.502	-7.992
9.00	-4.397	-5.371	-5.936	-6.482	-6.742	-7.045	-7.468	-7.909	-8.401	-8.953
10.00	-4.852	-5.933	-6.560	-7.168	-7.458	-7.794	-8.264	-8.755	-9.302	-9.918

Source: After Land, 1975.
This table is used in Section 13.2.

Source: Gilbert (1987).

Table B-9.
Quantiles of D'Agostino's Test for Normality



Quantiles of D'Agostino's Test for Normality (Values of Y Such That $100p\%$ of the Distribution of Y is Less Than Y_p)

n	$Y_{0.005}$	$Y_{0.01}$	$Y_{0.025}$	$Y_{0.05}$	$Y_{0.10}$	$Y_{0.90}$	$Y_{0.95}$	$Y_{0.975}$	$Y_{0.99}$	$Y_{0.995}$
50	-3.949	-3.442	-2.757	-2.220	-1.661	0.759	0.923	1.038	1.140	1.192
60	-3.846	-3.360	-2.699	-2.179	-1.634	0.807	0.986	1.115	1.236	1.301
70	-3.762	-3.293	-2.652	-2.146	-1.612	0.844	1.036	1.176	1.312	1.388
30	-3.693	-3.237	-2.613	-2.118	-1.594	0.874	1.076	1.226	1.374	1.459
90	-3.635	-3.100	-2.580	-2.095	-1.579	0.899	1.109	1.268	1.426	1.518
00	-3.584	-3.150	-2.552	-2.075	-1.566	0.920	1.137	1.303	1.470	1.569
50	-3.405	-3.009	-2.452	-2.004	-1.520	0.990	1.233	1.423	1.623	1.746
00	-3.302	-2.922	-2.391	-1.960	-1.491	1.032	1.290	1.496	1.715	1.853
50	-3.227	-2.861	-2.348	-1.926	-1.471	1.060	1.328	1.545	1.779	1.927
00	-3.172	-2.816	-2.316	-1.906	-1.456	1.080	1.357	1.528	1.826	1.963
50	-3.129	-2.781	-2.291	-1.888	-1.444	1.096	1.379	1.610	1.863	2.026
00	-3.094	-2.753	-2.270	-1.873	-1.434	1.108	1.396	1.633	1.893	2.061
50	-3.064	-2.729	-2.253	-1.861	-1.426	1.119	1.411	1.652	1.918	2.090
00	-3.040	-2.709	-2.239	-1.850	-1.419	1.127	1.423	1.668	1.938	2.114
50	-3.019	-2.691	-2.226	-1.841	-1.413	1.135	1.434	1.682	1.957	2.136
00	-3.000	-2.676	-2.215	-1.833	-1.408	1.141	1.443	1.694	1.972	2.154
0	-2.984	-2.663	-2.206	-1.826	-1.403	1.147	1.451	1.704	1.986	2.171
0	-2.969	-2.651	-2.197	-1.820	-1.399	1.152	1.458	1.714	1.999	2.185
0	-2.956	-2.640	-2.189	-1.814	-1.395	1.157	1.465	1.722	2.010	2.199
0	-2.944	-2.630	-2.182	-1.809	-1.392	1.161	1.471	1.730	2.020	2.211
850	-2.933	-2.621	-2.176	-1.804	-1.389	1.165	1.476	1.737	2.029	2.221
900	-2.923	-2.613	-2.170	-1.800	-1.386	1.168	1.481	1.743	2.037	2.231
950	-2.914	-2.605	-2.164	-1.796	-1.383	1.171	1.485	1.749	2.045	2.241
1000	-2.906	-2.599	-2.159	-1.792	-1.381	1.174	1.489	1.754	2.052	2.249

Source: From D'Agostino, 1971. Used by permission.

The null hypothesis of a normal distribution is rejected at the α significance level if the D'Agostino test statistic Y is less than $Y_{\alpha/2}$ or greater than $Y_{1-\alpha/2}$.

Source: Gilbert (1987).

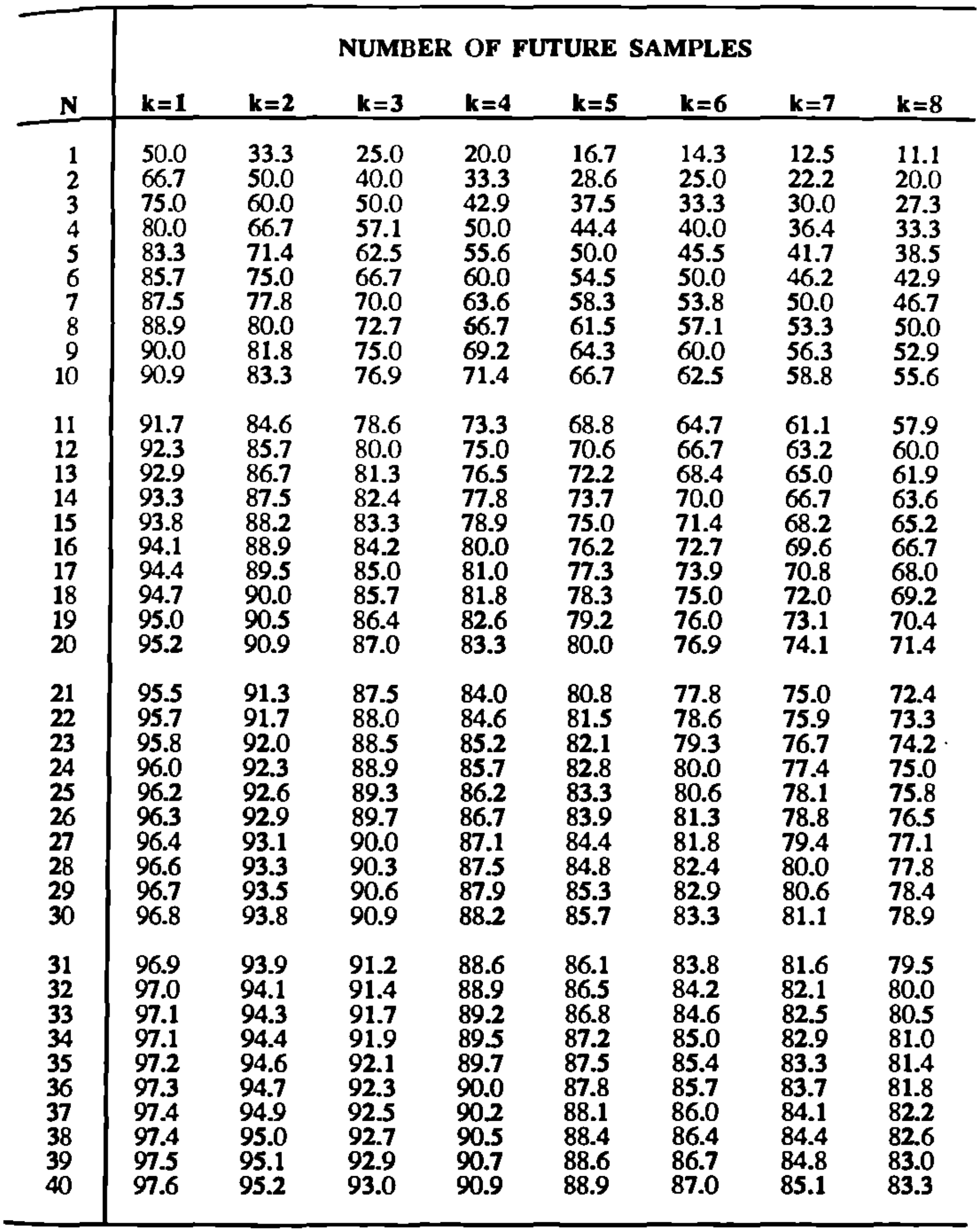
Table B-10.
Probabilities for the Small-Sample Mann-Kendall Test for Trend

S	n				S	n		
	4	5	8	9		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

S	***n***				**S**	***n***		
	4	5	8	9		6	7	10
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			0.000058
					41			0.000015
					43			0.0000028
					45			0.00000028

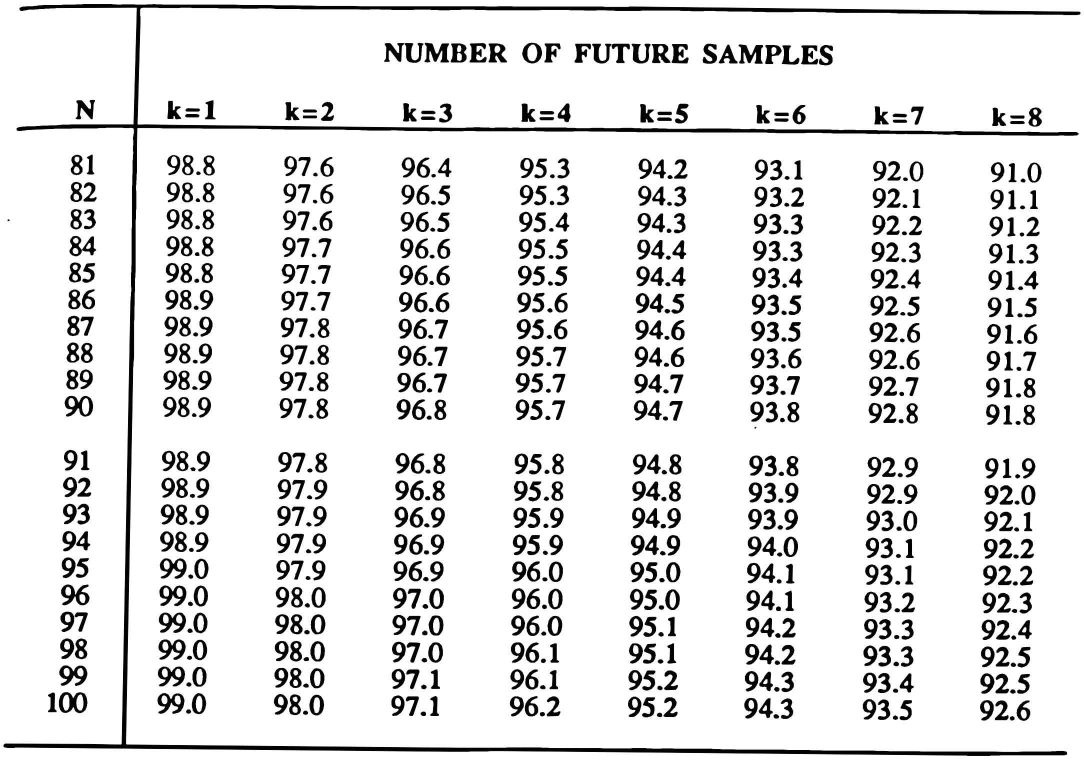
Source: EPA/600/R-96/084.

Table B-11.
Confidence Levels for Nonparametric Prediction Limits



	NUMBER OF FUTURE SAMPLES							
N	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8
1	50.0	33.3	25.0	20.0	16.7	14.3	12.5	11.1
2	66.7	50.0	40.0	33.3	28.6	25.0	22.2	20.0
3	75.0	60.0	50.0	42.9	37.5	33.3	30.0	27.3
4	80.0	66.7	57.1	50.0	44.4	40.0	36.4	33.3
5	83.3	71.4	62.5	55.6	50.0	45.5	41.7	38.5
6	85.7	75.0	66.7	60.0	54.5	50.0	46.2	42.9
7	87.5	77.8	70.0	63.6	58.3	53.8	50.0	46.7
8	88.9	80.0	72.7	66.7	61.5	57.1	53.3	50.0
9	90.0	81.8	75.0	69.2	64.3	60.0	56.3	52.9
10	90.9	83.3	76.9	71.4	66.7	62.5	58.8	55.6
11	91.7	84.6	78.6	73.3	68.8	64.7	61.1	57.9
12	92.3	85.7	80.0	75.0	70.6	66.7	63.2	60.0
13	92.9	86.7	81.3	76.5	72.2	68.4	65.0	61.9
14	93.3	87.5	82.4	77.8	73.7	70.0	66.7	63.6
15	93.8	88.2	83.3	78.9	75.0	71.4	68.2	65.2
16	94.1	88.9	84.2	80.0	76.2	72.7	69.6	66.7
17	94.4	89.5	85.0	81.0	77.3	73.9	70.8	68.0
18	94.7	90.0	85.7	81.8	78.3	75.0	72.0	69.2
19	95.0	90.5	86.4	82.6	79.2	76.0	73.1	70.4
20	95.2	90.9	87.0	83.3	80.0	76.9	74.1	71.4
21	95.5	91.3	87.5	84.0	80.8	77.8	75.0	72.4
22	95.7	91.7	88.0	84.6	81.5	78.6	75.9	73.3
23	95.8	92.0	88.5	85.2	82.1	79.3	76.7	74.2
24	96.0	92.3	88.9	85.7	82.8	80.0	77.4	75.0
25	96.2	92.6	89.3	86.2	83.3	80.6	78.1	75.8
26	96.3	92.9	89.7	86.7	83.9	81.3	78.8	76.5
27	96.4	93.1	90.0	87.1	84.4	81.8	79.4	77.1
28	96.6	93.3	90.3	87.5	84.8	82.4	80.0	77.8
29	96.7	93.5	90.6	87.9	85.3	82.9	80.6	78.4
30	96.8	93.8	90.9	88.2	85.7	83.3	81.1	78.9
31	96.9	93.9	91.2	88.6	86.1	83.8	81.6	79.5
32	97.0	94.1	91.4	88.9	86.5	84.2	82.1	80.0
33	97.1	94.3	91.7	89.2	86.8	84.6	82.5	80.5
34	97.1	94.4	91.9	89.5	87.2	85.0	82.9	81.0
35	97.2	94.6	92.1	89.7	87.5	85.4	83.3	81.4
36	97.3	94.7	92.3	90.0	87.8	85.7	83.7	81.8
37	97.4	94.9	92.5	90.2	88.1	86.0	84.1	82.2
38	97.4	95.0	92.7	90.5	88.4	86.4	84.4	82.6
39	97.5	95.1	92.9	90.7	88.6	86.7	84.8	83.0
40	97.6	95.2	93.0	90.9	88.9	87.0	85.1	83.3

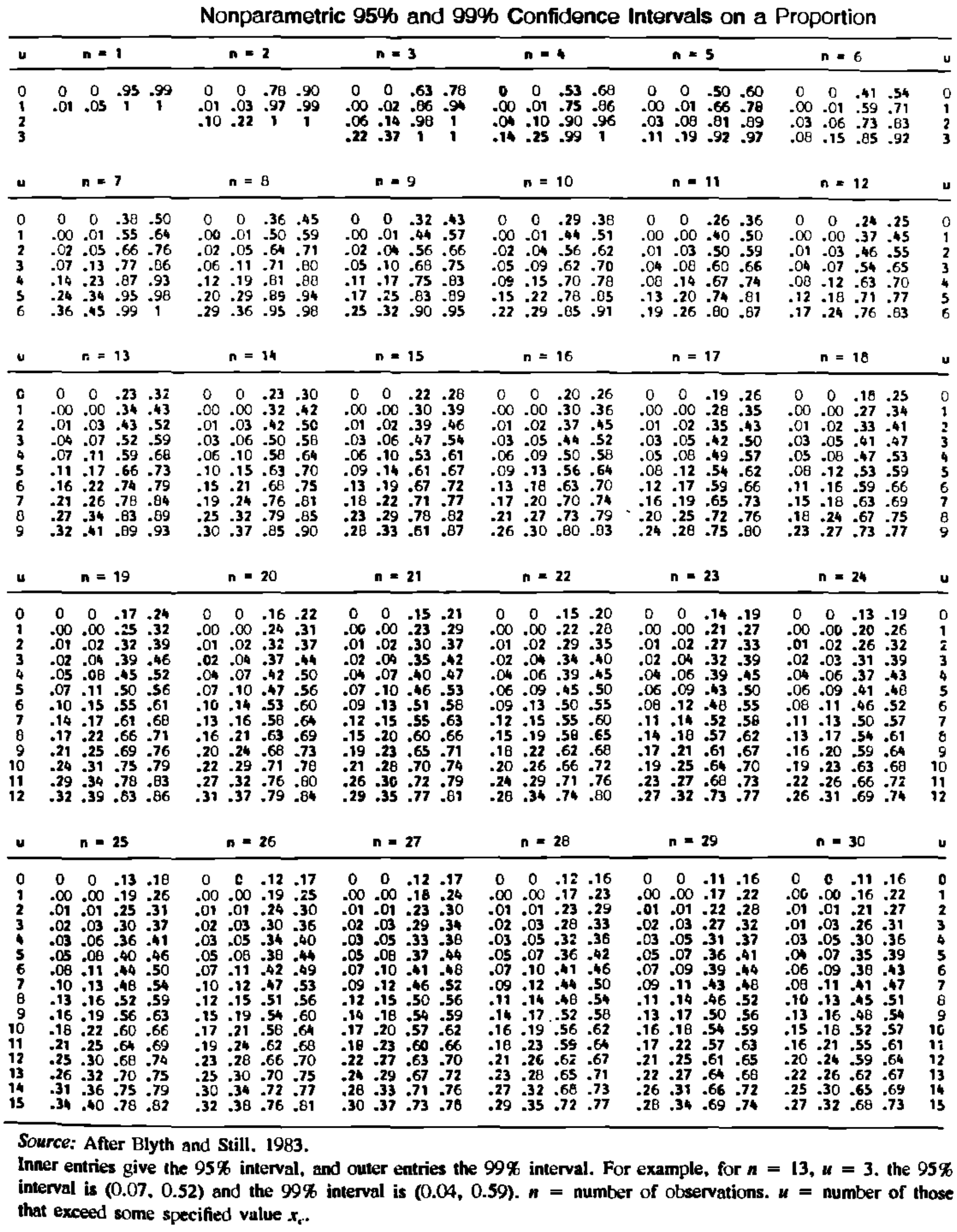
	NUMBER OF FUTURE SAMPLES							
N	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8
41	97.6	95.3	93.2	91.1	89.1	87.2	85.4	83.7
42	97.7	95.5	93.3	91.3	89.4	87.5	85.7	84.0
43	97.7	95.6	93.5	91.5	89.6	87.8	86.0	84.3
44	97.8	95.7	93.6	91.7	89.8	88.0	86.3	84.6
45	97.8	95.7	93.8	91.8	90.0	88.2	86.5	84.9
46	97.9	95.8	93.9	92.0	90.2	88.5	86.8	85.2
47	97.9	95.9	94.0	92.2	90.4	88.7	87.0	85.5
48	98.0	96.0	94.1	92.3	90.6	88.9	87.3	85.7
49	98.0	96.1	94.2	92.5	90.7	89.1	87.5	86.0
50	98.0	96.2	94.3	92.6	90.9	89.3	87.7	86.2
51	98.1	96.2	94.4	92.7	91.1	89.5	87.9	86.4
52	98.1	96.3	94.5	92.9	91.2	89.7	88.1	86.7
53	98.1	96.4	94.6	93.0	91.4	89.8	88.3	86.9
54	98.2	96.4	94.7	93.1	91.5	90.0	88.5	87.1
55	98.2	96.5	94.8	93.2	91.7	90.2	88.7	87.3
56	98.2	96.6	94.9	93.3	91.8	90.3	88.9	87.5
57	98.3	96.6	95.0	93.4	91.9	90.5	89.1	87.7
58	98.3	96.7	95.1	93.5	92.1	90.6	89.2	87.9
59	98.3	96.7	95.2	93.7	92.2	90.8	89.4	88.1
60	98.4	96.8	95.2	93.8	92.3	90.9	89.6	88.2
61	98.4	96.8	95.3	93.8	92.4	91.0	89.7	88.4
62	98.4	96.9	95.4	93.9	92.5	91.2	89.9	88.6
63	98.4	96.9	95.5	94.0	92.6	91.3	90.0	88.7
64	98.5	97.0	95.5	94.1	92.8	91.4	90.1	88.9
65	98.5	97.0	95.6	94.2	92.9	91.5	90.3	89.0
66	98.5	97.1	95.7	94.3	93.0	91.7	90.4	89.2
67	98.5	97.1	95.7	94.4	93.1	91.8	90.5	89.3
68	98.6	97.1	95.8	94.4	93.2	91.9	90.7	89.5
69	98.6	97.2	95.8	94.5	93.2	92.0	90.8	89.6
70	98.6	97.2	95.9	94.6	93.3	92.1	90.9	89.7
71	98.6	97.3	95.9	94.7	93.4	92.2	91.0	89.9
72	98.6	97.3	96.0	94.7	93.5	92.3	91.1	90.0
73	98.6	97.3	96.1	94.8	93.6	92.4	91.3	90.1
74	98.7	97.4	96.1	94.9	93.7	92.5	91.4	90.2
75	98.7	97.4	96.2	94.9	93.8	92.6	91.5	90.4
76	98.7	97.4	96.2	95.0	93.8	92.7	91.6	90.5
77	98.7	97.5	96.3	95.1	93.9	92.8	91.7	90.6
78	98.7	97.5	96.3	95.1	94.0	92.9	91.8	90.7
79	98.8	97.5	96.3	95.2	94.0	92.9	91.9	90.8
80	98.8	97.6	96.4	95.2	94.1	93.0	92.0	90.9



	NUMBER OF FUTURE SAMPLES							
N	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8
81	98.8	97.6	96.4	95.3	94.2	93.1	92.0	91.0
82	98.8	97.6	96.5	95.3	94.3	93.2	92.1	91.1
83	98.8	97.6	96.5	95.4	94.3	93.3	92.2	91.2
84	98.8	97.7	96.6	95.5	94.4	93.3	92.3	91.3
85	98.8	97.7	96.6	95.5	94.4	93.4	92.4	91.4
86	98.9	97.7	96.6	95.6	94.5	93.5	92.5	91.5
87	98.9	97.8	96.7	95.6	94.6	93.5	92.6	91.6
88	98.9	97.8	96.7	95.7	94.6	93.6	92.6	91.7
89	98.9	97.8	96.7	95.7	94.7	93.7	92.7	91.8
90	98.9	97.8	96.8	95.7	94.7	93.8	92.8	91.8
91	98.9	97.8	96.8	95.8	94.8	93.8	92.9	91.9
92	98.9	97.9	96.8	95.8	94.8	93.9	92.9	92.0
93	98.9	97.9	96.9	95.9	94.9	93.9	93.0	92.1
94	98.9	97.9	96.9	95.9	94.9	94.0	93.1	92.2
95	99.0	97.9	96.9	96.0	95.0	94.1	93.1	92.2
96	99.0	98.0	97.0	96.0	95.0	94.1	93.2	92.3
97	99.0	98.0	97.0	96.0	95.1	94.2	93.3	92.4
98	99.0	98.0	97.0	96.1	95.1	94.2	93.3	92.5
99	99.0	98.0	97.1	96.1	95.2	94.3	93.4	92.5
100	99.0	98.0	97.1	96.2	95.2	94.3	93.5	92.6

Source: EPA/530-SW-89-026.

Table B-12.
Nonparametric Confidence Intervals on a Proportion



Nonparametric 95% and 99% Confidence Intervals on a Proportion

u	n = 1				n = 2				n = 3				n = 4				n = 5				n = 6				u
0	0	0	.95	.99	0	0	.78	.90	0	0	.63	.78	0	0	.53	.68	0	0	.50	.60	0	0	.41	.54	0
1	.01	.05	1	1	.01	.03	.97	.99	.00	.02	.86	.94	.00	.01	.75	.86	.00	.01	.66	.78	.00	.01	.59	.71	1
2					.10	.22	1	1	.06	.14	.98	1	.04	.10	.90	.96	.03	.08	.81	.89	.03	.06	.73	.83	2
3									.22	.37	1	1	.14	.25	.99	1	.11	.19	.92	.97	.08	.15	.85	.92	3

u	n = 7				n = 8				n = 9				n = 10				n = 11				n = 12				u
0	0	0	.38	.50	0	0	.36	.45	0	0	.32	.43	0	0	.29	.38	0	0	.26	.36	0	0	.24	.25	0
1	.00	.01	.55	.64	.00	.01	.50	.59	.00	.01	.44	.57	.00	.01	.44	.51	.00	.00	.40	.50	.00	.00	.37	.45	1
2	.02	.05	.66	.76	.02	.05	.64	.71	.02	.04	.56	.66	.02	.04	.56	.62	.01	.03	.50	.59	.01	.03	.46	.55	2
3	.07	.13	.77	.86	.06	.11	.71	.80	.05	.10	.68	.75	.05	.09	.62	.70	.04	.08	.60	.66	.04	.07	.54	.65	3
4	.14	.23	.87	.93	.12	.19	.81	.88	.11	.17	.75	.83	.09	.15	.70	.78	.08	.14	.67	.74	.08	.12	.63	.70	4
5	.24	.34	.95	.98	.20	.29	.89	.94	.17	.25	.83	.89	.15	.22	.78	.85	.13	.20	.74	.81	.12	.18	.71	.77	5
6	.36	.45	.99	1	.29	.36	.95	.98	.25	.32	.90	.95	.22	.29	.85	.91	.19	.26	.80	.87	.17	.24	.76	.83	6

u	n = 13				n = 14				n = 15				n = 16				n = 17				n = 18				u
0	0	0	.23	.32	0	0	.23	.30	0	0	.22	.28	0	0	.20	.26	0	0	.19	.26	0	0	.18	.25	0
1	.00	.00	.34	.43	.00	.00	.32	.42	.00	.00	.30	.39	.00	.00	.30	.36	.00	.00	.28	.35	.00	.00	.27	.34	1
2	.01	.03	.43	.52	.01	.03	.42	.50	.01	.02	.39	.46	.01	.02	.37	.45	.01	.02	.35	.43	.01	.02	.33	.41	2
3	.04	.07	.52	.59	.03	.06	.50	.58	.03	.06	.47	.54	.03	.05	.44	.52	.03	.05	.42	.50	.03	.05	.41	.47	3
4	.07	.11	.59	.68	.06	.10	.58	.64	.06	.10	.53	.61	.06	.09	.50	.58	.05	.08	.49	.57	.05	.08	.47	.53	4
5	.11	.17	.66	.73	.10	.15	.63	.70	.09	.14	.61	.67	.09	.13	.56	.64	.08	.12	.54	.62	.08	.12	.53	.59	5
6	.16	.22	.74	.79	.15	.21	.68	.75	.13	.19	.67	.72	.13	.18	.63	.70	.12	.17	.59	.66	.11	.16	.59	.66	6
7	.21	.26	.78	.84	.19	.24	.76	.81	.18	.22	.71	.77	.17	.20	.70	.74	.16	.19	.65	.73	.15	.18	.63	.69	7
8	.27	.34	.83	.89	.25	.32	.79	.85	.23	.29	.78	.82	.21	.27	.73	.79	.20	.25	.72	.76	.18	.24	.67	.75	8
9	.32	.41	.89	.93	.30	.37	.85	.90	.28	.33	.61	.87	.26	.30	.80	.83	.24	.28	.75	.80	.23	.27	.73	.77	9

u	n = 19				n = 20				n = 21				n = 22				n = 23				n = 24				u
0	0	0	.17	.24	0	0	.16	.22	0	0	.15	.21	0	0	.15	.20	0	0	.14	.19	0	0	.13	.19	0
1	.00	.00	.25	.32	.00	.00	.24	.31	.00	.00	.23	.29	.00	.00	.22	.28	.00	.00	.21	.27	.00	.00	.20	.26	1
2	.01	.02	.32	.39	.01	.02	.32	.37	.01	.02	.30	.37	.01	.02	.29	.35	.01	.02	.27	.33	.01	.02	.26	.32	2
3	.02	.04	.39	.46	.02	.04	.37	.44	.02	.04	.35	.42	.02	.04	.34	.40	.02	.04	.32	.39	.02	.03	.31	.39	3
4	.05	.08	.45	.52	.04	.07	.42	.50	.04	.07	.40	.47	.04	.06	.39	.45	.04	.06	.39	.45	.04	.06	.37	.43	4
5	.07	.11	.50	.56	.07	.10	.47	.56	.07	.10	.46	.53	.06	.09	.45	.50	.06	.09	.43	.50	.06	.09	.41	.48	5
6	.10	.15	.55	.61	.10	.14	.53	.60	.09	.13	.51	.58	.09	.13	.50	.55	.08	.12	.48	.55	.08	.11	.46	.52	6
7	.14	.17	.61	.68	.13	.16	.58	.64	.12	.15	.55	.63	.12	.15	.55	.60	.11	.14	.52	.58	.11	.13	.50	.57	7
8	.17	.22	.66	.71	.16	.21	.63	.69	.15	.20	.60	.66	.15	.19	.58	.65	.14	.18	.57	.62	.13	.17	.54	.61	8
9	.21	.25	.69	.76	.20	.24	.68	.73	.19	.23	.65	.71	.18	.22	.62	.68	.17	.21	.61	.67	.16	.20	.59	.64	9
10	.24	.31	.75	.79	.22	.29	.71	.78	.21	.28	.70	.74	.20	.26	.66	.72	.19	.25	.64	.70	.19	.23	.63	.68	10
11	.29	.34	.78	.83	.27	.32	.76	.80	.26	.30	.72	.79	.24	.29	.71	.76	.23	.27	.68	.73	.22	.26	.66	.72	11
12	.32	.39	.83	.86	.31	.37	.79	.84	.29	.35	.77	.81	.28	.34	.74	.80	.27	.32	.73	.77	.26	.31	.69	.74	12

u	n = 25				n = 26				n = 27				n = 28				n = 29				n = 30				u
0	0	0	.13	.18	0	0	.12	.17	0	0	.12	.17	0	0	.12	.16	0	0	.11	.16	0	0	.11	.16	0
1	.00	.00	.19	.26	.00	.00	.19	.25	.00	.00	.18	.24	.00	.00	.17	.23	.00	.00	.17	.22	.00	.00	.16	.22	1
2	.01	.01	.25	.31	.01	.01	.24	.30	.01	.01	.23	.30	.01	.01	.23	.29	.01	.01	.22	.28	.01	.01	.21	.27	2
3	.02	.03	.30	.37	.02	.03	.30	.36	.02	.03	.29	.34	.02	.03	.28	.33	.02	.03	.27	.32	.01	.03	.26	.31	3
4	.03	.06	.36	.41	.03	.05	.34	.40	.03	.05	.33	.38	.03	.05	.32	.36	.03	.05	.31	.37	.03	.05	.30	.36	4
5	.05	.08	.40	.46	.05	.08	.38	.44	.05	.08	.37	.44	.05	.07	.36	.42	.05	.07	.36	.41	.04	.07	.35	.39	5
6	.08	.11	.44	.50	.07	.11	.42	.49	.07	.10	.41	.48	.07	.10	.41	.46	.07	.09	.39	.44	.06	.09	.38	.43	6
7	.10	.13	.48	.54	.10	.12	.47	.53	.09	.12	.46	.52	.09	.12	.44	.50	.09	.11	.43	.48	.08	.11	.41	.47	7
8	.13	.16	.52	.59	.12	.15	.51	.56	.12	.15	.50	.56	.11	.14	.48	.54	.11	.14	.46	.52	.10	.13	.45	.51	8
9	.16	.19	.56	.63	.15	.19	.54	.60	.14	.18	.54	.59	.14	.17	.52	.58	.13	.17	.50	.56	.13	.16	.48	.54	9
10	.18	.22	.60	.66	.17	.21	.58	.64	.17	.20	.57	.62	.16	.19	.56	.62	.16	.18	.54	.59	.15	.18	.52	.57	10
11	.21	.25	.64	.69	.19	.24	.62	.68	.18	.23	.60	.66	.18	.23	.59	.64	.17	.22	.57	.63	.16	.21	.55	.61	11
12	.25	.30	.68	.74	.23	.28	.66	.70	.22	.27	.63	.70	.21	.26	.62	.67	.21	.25	.61	.65	.20	.24	.59	.64	12
13	.26	.32	.70	.75	.25	.30	.70	.75	.24	.29	.67	.72	.23	.28	.65	.71	.22	.27	.64	.68	.22	.26	.62	.67	13
14	.31	.36	.75	.79	.30	.34	.72	.77	.28	.33	.71	.76	.27	.32	.68	.73	.26	.31	.66	.72	.25	.30	.65	.69	14
15	.34	.40	.78	.82	.32	.38	.76	.81	.30	.37	.73	.78	.29	.35	.72	.77	.28	.34	.69	.74	.27	.32	.68	.73	15

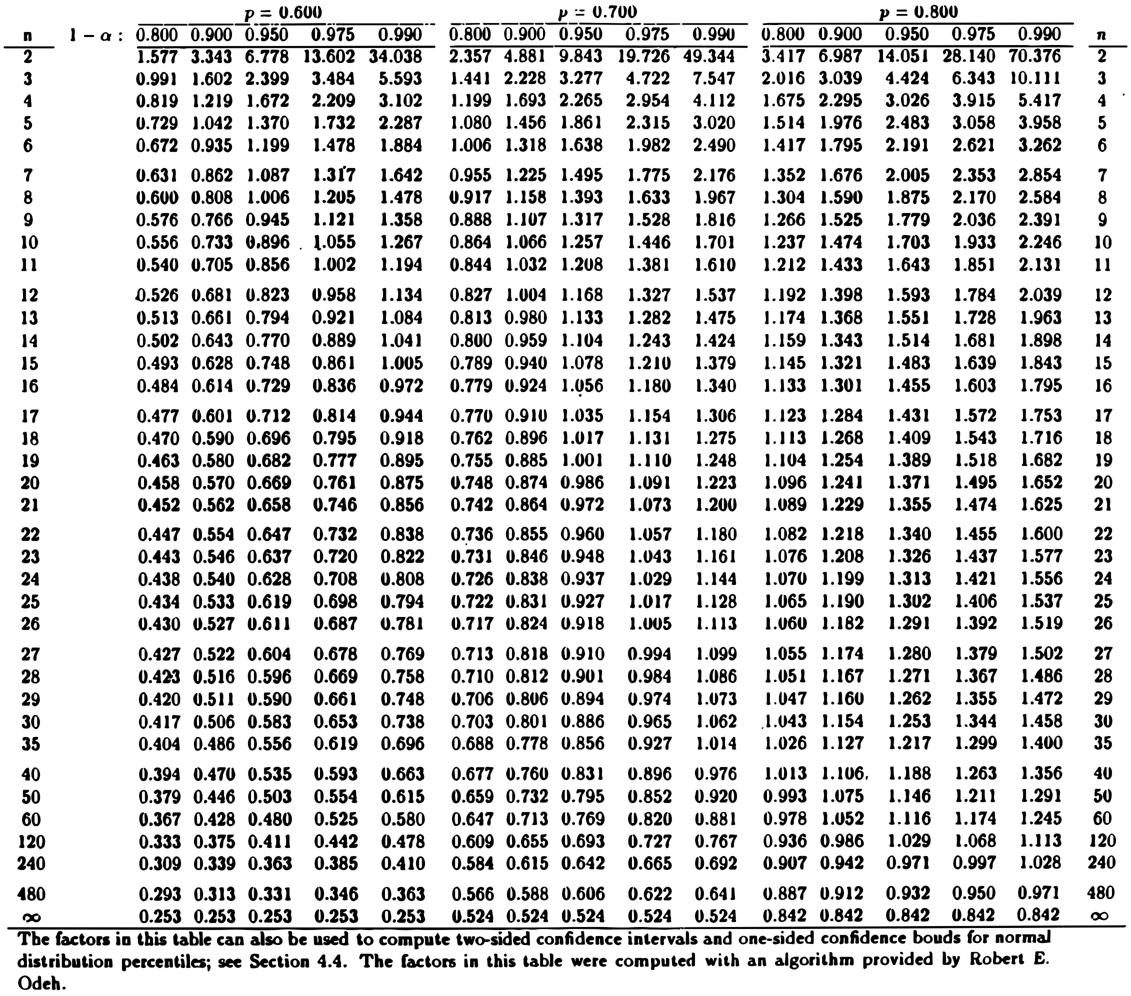
Source: After Blyth and Still, 1983.

Inner entries give the 95% interval, and outer entries the 99% interval. For example, for *n* = 13, *u* = 3, the 95% interval is (0.07, 0.52) and the 99% interval is (0.04, 0.59). *n* = number of observations. *u* = number of those that exceed some specified value x_c.

Source: Gilbert (1987).

Table B-13.
Factors for Calculating Normal Distribution One-Sided Tolerance Limits

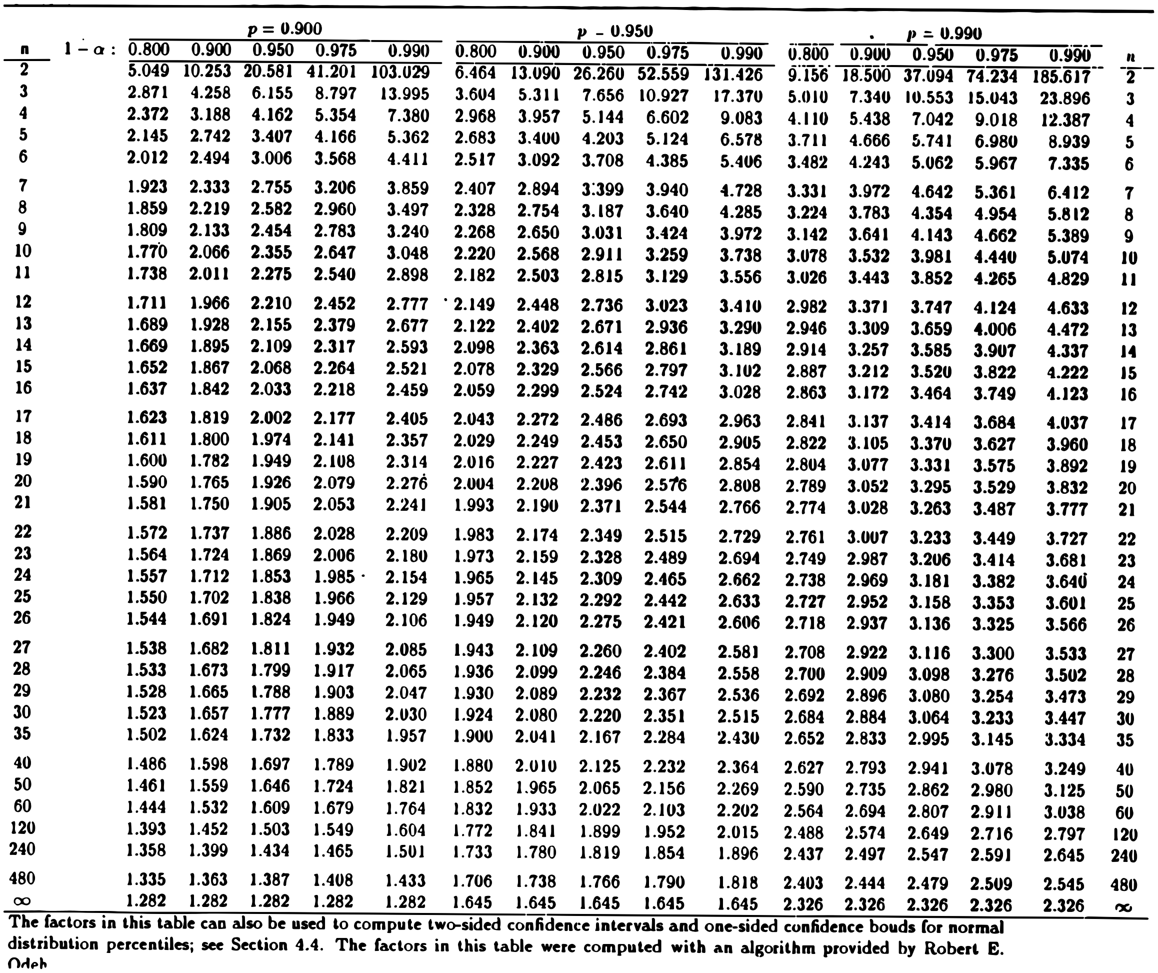
Factors $g'_{(1-\alpha,p,n)}$ for a Normal One-Sided $(1-\alpha)$100% Tolerance Bound



		$p = 0.600$					$p = 0.700$					$p = 0.800$					
n	1 − α :	0.800	0.900	0.950	0.975	0.990	0.800	0.900	0.950	0.975	0.990	0.800	0.900	0.950	0.975	0.990	n
2		1.577	3.343	6.778	13.602	34.038	2.357	4.881	9.843	19.726	49.344	3.417	6.987	14.051	28.140	70.376	2
3		0.991	1.602	2.399	3.484	5.593	1.441	2.228	3.277	4.722	7.547	2.016	3.039	4.424	6.343	10.111	3
4		0.819	1.219	1.672	2.209	3.102	1.199	1.693	2.265	2.954	4.112	1.675	2.295	3.026	3.915	5.417	4
5		0.729	1.042	1.370	1.732	2.287	1.080	1.456	1.861	2.315	3.020	1.514	1.976	2.483	3.058	3.958	5
6		0.672	0.935	1.199	1.478	1.884	1.006	1.318	1.638	1.982	2.490	1.417	1.795	2.191	2.621	3.262	6
7		0.631	0.862	1.087	1.317	1.642	0.955	1.225	1.495	1.775	2.176	1.352	1.676	2.005	2.353	2.854	7
8		0.600	0.808	1.006	1.205	1.478	0.917	1.158	1.393	1.633	1.967	1.304	1.590	1.875	2.170	2.584	8
9		0.576	0.766	0.945	1.121	1.358	0.888	1.107	1.317	1.528	1.816	1.266	1.525	1.779	2.036	2.391	9
10		0.556	0.733	0.896	1.055	1.267	0.864	1.066	1.257	1.446	1.701	1.237	1.474	1.703	1.933	2.246	10
11		0.540	0.705	0.856	1.002	1.194	0.844	1.032	1.208	1.381	1.610	1.212	1.433	1.643	1.851	2.131	11
12		0.526	0.681	0.823	0.958	1.134	0.827	1.004	1.168	1.327	1.537	1.192	1.398	1.593	1.784	2.039	12
13		0.513	0.661	0.794	0.921	1.084	0.813	0.980	1.133	1.282	1.475	1.174	1.368	1.551	1.728	1.963	13
14		0.502	0.643	0.770	0.889	1.041	0.800	0.959	1.104	1.243	1.424	1.159	1.343	1.514	1.681	1.898	14
15		0.493	0.628	0.748	0.861	1.005	0.789	0.940	1.078	1.210	1.379	1.145	1.321	1.483	1.639	1.843	15
16		0.484	0.614	0.729	0.836	0.972	0.779	0.924	1.056	1.180	1.340	1.133	1.301	1.455	1.603	1.795	16
17		0.477	0.601	0.712	0.814	0.944	0.770	0.910	1.035	1.154	1.306	1.123	1.284	1.431	1.572	1.753	17
18		0.470	0.590	0.696	0.795	0.918	0.762	0.896	1.017	1.131	1.275	1.113	1.268	1.409	1.543	1.716	18
19		0.463	0.580	0.682	0.777	0.895	0.755	0.885	1.001	1.110	1.248	1.104	1.254	1.389	1.518	1.682	19
20		0.458	0.570	0.669	0.761	0.875	0.748	0.874	0.986	1.091	1.223	1.096	1.241	1.371	1.495	1.652	20
21		0.452	0.562	0.658	0.746	0.856	0.742	0.864	0.972	1.073	1.200	1.089	1.229	1.355	1.474	1.625	21
22		0.447	0.554	0.647	0.732	0.838	0.736	0.855	0.960	1.057	1.180	1.082	1.218	1.340	1.455	1.600	22
23		0.443	0.546	0.637	0.720	0.822	0.731	0.846	0.948	1.043	1.161	1.076	1.208	1.326	1.437	1.577	23
24		0.438	0.540	0.628	0.708	0.808	0.726	0.838	0.937	1.029	1.144	1.070	1.199	1.313	1.421	1.556	24
25		0.434	0.533	0.619	0.698	0.794	0.722	0.831	0.927	1.017	1.128	1.065	1.190	1.302	1.406	1.537	25
26		0.430	0.527	0.611	0.687	0.781	0.717	0.824	0.918	1.005	1.113	1.060	1.182	1.291	1.392	1.519	26
27		0.427	0.522	0.604	0.678	0.769	0.713	0.818	0.910	0.994	1.099	1.055	1.174	1.280	1.379	1.502	27
28		0.423	0.516	0.596	0.669	0.758	0.710	0.812	0.901	0.984	1.086	1.051	1.167	1.271	1.367	1.486	28
29		0.420	0.511	0.590	0.661	0.748	0.706	0.806	0.894	0.974	1.073	1.047	1.160	1.262	1.355	1.472	29
30		0.417	0.506	0.583	0.653	0.738	0.703	0.801	0.886	0.965	1.062	1.043	1.154	1.253	1.344	1.458	30
35		0.404	0.486	0.556	0.619	0.696	0.688	0.778	0.856	0.927	1.014	1.026	1.127	1.217	1.299	1.400	35
40		0.394	0.470	0.535	0.593	0.663	0.677	0.760	0.831	0.896	0.976	1.013	1.106	1.188	1.263	1.356	40
50		0.379	0.446	0.503	0.554	0.615	0.659	0.732	0.795	0.852	0.920	0.993	1.075	1.146	1.211	1.291	50
60		0.367	0.428	0.480	0.525	0.580	0.647	0.713	0.769	0.820	0.881	0.978	1.052	1.116	1.174	1.245	60
120		0.333	0.375	0.411	0.442	0.478	0.609	0.655	0.693	0.727	0.767	0.936	0.986	1.029	1.068	1.113	120
240		0.309	0.339	0.363	0.385	0.410	0.584	0.615	0.642	0.665	0.692	0.907	0.942	0.971	0.997	1.028	240
480		0.293	0.313	0.331	0.346	0.363	0.566	0.588	0.606	0.622	0.641	0.887	0.912	0.932	0.950	0.971	480
∞		0.253	0.253	0.253	0.253	0.253	0.524	0.524	0.524	0.524	0.524	0.842	0.842	0.842	0.842	0.842	∞

The factors in this table can also be used to compute two-sided confidence intervals and one-sided confidence bouds for normal distribution percentiles; see Section 4.4. The factors in this table were computed with an algorithm provided by Robert E. Odeh.

Factors $g'_{(1-\alpha,p,n)}$ for a Normal One-Sided $(1-\alpha)$100% Tolerance Bound



		$p = 0.900$					p - 0.950					$p = 0.990$					
n	$1-\alpha$:	0.800	0.900	0.950	0.975	0.990	0.800	0.900	0.950	0.975	0.990	0.800	0.900	0.950	0.975	0.990	n
2		5.049	10.253	20.581	41.201	103.029	6.464	13.090	26.260	52.559	131.426	9.156	18.500	37.094	74.234	185.617	2
3		2.871	4.258	6.155	8.797	13.995	3.604	5.311	7.656	10.927	17.370	5.010	7.340	10.553	15.043	23.896	3
4		2.372	3.188	4.162	5.354	7.380	2.968	3.957	5.144	6.602	9.083	4.110	5.438	7.042	9.018	12.387	4
5		2.145	2.742	3.407	4.166	5.362	2.683	3.400	4.203	5.124	6.578	3.711	4.666	5.741	6.980	8.939	5
6		2.012	2.494	3.006	3.568	4.411	2.517	3.092	3.708	4.385	5.406	3.482	4.243	5.062	5.967	7.335	6
7		1.923	2.333	2.755	3.206	3.859	2.407	2.894	3.399	3.940	4.728	3.331	3.972	4.642	5.361	6.412	7
8		1.859	2.219	2.582	2.960	3.497	2.328	2.754	3.187	3.640	4.285	3.224	3.783	4.354	4.954	5.812	8
9		1.809	2.133	2.454	2.783	3.240	2.268	2.650	3.031	3.424	3.972	3.142	3.641	4.143	4.662	5.389	9
10		1.770	2.066	2.355	2.647	3.048	2.220	2.568	2.911	3.259	3.738	3.078	3.532	3.981	4.440	5.074	10
11		1.738	2.011	2.275	2.540	2.898	2.182	2.503	2.815	3.129	3.556	3.026	3.443	3.852	4.265	4.829	11
12		1.711	1.966	2.210	2.452	2.777	2.149	2.448	2.736	3.023	3.410	2.982	3.371	3.747	4.124	4.633	12
13		1.689	1.928	2.155	2.379	2.677	2.122	2.402	2.671	2.936	3.290	2.946	3.309	3.659	4.006	4.472	13
14		1.669	1.895	2.109	2.317	2.593	2.098	2.363	2.614	2.861	3.189	2.914	3.257	3.585	3.907	4.337	14
15		1.652	1.867	2.068	2.264	2.521	2.078	2.329	2.566	2.797	3.102	2.887	3.212	3.520	3.822	4.222	15
16		1.637	1.842	2.033	2.218	2.459	2.059	2.299	2.524	2.742	3.028	2.863	3.172	3.464	3.749	4.123	16
17		1.623	1.819	2.002	2.177	2.405	2.043	2.272	2.486	2.693	2.963	2.841	3.137	3.414	3.684	4.037	17
18		1.611	1.800	1.974	2.141	2.357	2.029	2.249	2.453	2.650	2.905	2.822	3.105	3.370	3.627	3.960	18
19		1.600	1.782	1.949	2.108	2.314	2.016	2.227	2.423	2.611	2.854	2.804	3.077	3.331	3.575	3.892	19
20		1.590	1.765	1.926	2.079	2.276	2.004	2.208	2.396	2.576	2.808	2.789	3.052	3.295	3.529	3.832	20
21		1.581	1.750	1.905	2.053	2.241	1.993	2.190	2.371	2.544	2.766	2.774	3.028	3.263	3.487	3.777	21
22		1.572	1.737	1.886	2.028	2.209	1.983	2.174	2.349	2.515	2.729	2.761	3.007	3.233	3.449	3.727	22
23		1.564	1.724	1.869	2.006	2.180	1.973	2.159	2.328	2.489	2.694	2.749	2.987	3.206	3.414	3.681	23
24		1.557	1.712	1.853	1.985	2.154	1.965	2.145	2.309	2.465	2.662	2.738	2.969	3.181	3.382	3.640	24
25		1.550	1.702	1.838	1.966	2.129	1.957	2.132	2.292	2.442	2.633	2.727	2.952	3.158	3.353	3.601	25
26		1.544	1.691	1.824	1.949	2.106	1.949	2.120	2.275	2.421	2.606	2.718	2.937	3.136	3.325	3.566	26
27		1.538	1.682	1.811	1.932	2.085	1.943	2.109	2.260	2.402	2.581	2.708	2.922	3.116	3.300	3.533	27
28		1.533	1.673	1.799	1.917	2.065	1.936	2.099	2.246	2.384	2.558	2.700	2.909	3.098	3.276	3.502	28
29		1.528	1.665	1.788	1.903	2.047	1.930	2.089	2.232	2.367	2.536	2.692	2.896	3.080	3.254	3.473	29
30		1.523	1.657	1.777	1.889	2.030	1.924	2.080	2.220	2.351	2.515	2.684	2.884	3.064	3.233	3.447	30
35		1.502	1.624	1.732	1.833	1.957	1.900	2.041	2.167	2.284	2.430	2.652	2.833	2.995	3.145	3.334	35
40		1.486	1.598	1.697	1.789	1.902	1.880	2.010	2.125	2.232	2.364	2.627	2.793	2.941	3.078	3.249	40
50		1.461	1.559	1.646	1.724	1.821	1.852	1.965	2.065	2.156	2.269	2.590	2.735	2.862	2.980	3.125	50
60		1.444	1.532	1.609	1.679	1.764	1.832	1.933	2.022	2.103	2.202	2.564	2.694	2.807	2.911	3.038	60
120		1.393	1.452	1.503	1.549	1.604	1.772	1.841	1.899	1.952	2.015	2.488	2.574	2.649	2.716	2.797	120
240		1.358	1.399	1.434	1.465	1.501	1.733	1.780	1.819	1.854	1.896	2.437	2.497	2.547	2.591	2.645	240
480		1.335	1.363	1.387	1.408	1.433	1.706	1.738	1.766	1.790	1.818	2.403	2.444	2.479	2.509	2.545	480
∞		1.282	1.282	1.282	1.282	1.282	1.645	1.645	1.645	1.645	1.645	2.326	2.326	2.326	2.326	2.326	∞

The factors in this table can also be used to compute two-sided confidence intervals and one-sided confidence bouds for normal distribution percentiles; see Section 4.4. The factors in this table were computed with an algorithm provided by Robert E. Odeh

Source: Hahn and Meeker (1991).

Table B-14.
Factors for Calculating Normal Distribution Two-Sided Tolerance Intervals

Factors $g_{(1-\alpha,p,n)}$ for a Normal Two-Sided $(1-\alpha)100\%$ Tolerance Intervals to Contain at Least $p100\%$ of the Population

	p = 0.90			*p* = 0.95			*p* = 0.99		
1 - α	0.90	0.95	0.99	0.9	0.95	0.99	0.90	0.95	0.99
n									
10	2.535	2.838	3.582	3.021	3.3819	4.268	3.970	4.445	5.609
11	2.463	2.737	3.397	2.935	3.2612	4.047	3.857	4.286	5.319
12	2.404	2.655	3.249	2.865	3.1633	3.872	3.765	4.157	5.089
13	2.355	2.587	3.129	2.806	3.0821	3.729	3.688	4.051	4.900
14	2.313	2.529	3.029	2.757	3.0135	3.610	3.623	3.960	4.744
15	2.278	2.480	2.944	2.714	2.9548	3.508	3.567	3.883	4.611
16	2.246	2.437	2.872	2.676	2.9038	3.422	3.517	3.816	4.497
17	2.219	2.399	2.808	2.644	2.859	3.346	3.474	3.757	4.398
18	2.194	2.366	2.753	2.614	2.8194	3.280	3.436	3.705	4.311
19	2.172	2.337	2.703	2.588	2.7841	3.221	3.402	3.659	4.233
20	2.152	2.310	2.659	2.565	2.7523	3.169	3.371	3.617	4.164
25	2.077	2.208	2.494	2.475	2.6313	2.972	3.252	3.458	3.906
30	2.025	2.140	2.385	2.413	2.5496	2.842	3.171	3.351	3.735
35	1.988	2.090	2.306	2.368	2.4902	2.748	3.112	3.273	3.612
40	1.959	2.052	2.247	2.334	2.4446	2.677	3.067	3.213	3.518
50	1.916	1.996	2.162	2.284	2.3788	2.576	3.001	3.126	3.385
60	1.887	1.958	2.103	2.249	2.3329	2.506	2.955	3.066	3.293
70	1.865	1.929	2.060	2.222	2.2987	2.454	2.921	3.021	3.225
80	1.848	1.907	2.026	2.202	2.2721	2.414	2.894	2.986	3.173
90	1.834	1.889	1.999	2.185	2.2506	2.382	2.872	2.958	3.130
100	1.823	1.874	1.977	2.172	2.2328	2.356	2.854	2.934	3.096
∞	1.645	1.645	1.645	1.960	1.960	1.960	2.576	2.576	2.576

Table B-15.
Standard Normal Distribution

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003369	0.0003248	0.0003131	0.0003018	0.0002909	0.0002803	0.0002701	0.0002602	0.0002507	0.0002415
-3.3	0.0004834	0.0004665	0.0004501	0.0004342	0.0004189	0.0004041	0.0003897	0.0003758	0.0003624	0.0003495
-3.2	0.0006871	0.0006637	0.0006410	0.0006190	0.0005976	0.0005770	0.0005571	0.0005377	0.0005190	0.0005009
-3.1	0.0009676	0.0009354	0.0009043	0.0008740	0.0008447	0.0008164	0.0007888	0.0007622	0.0007364	0.0007114
-3.0	0.001350	0.001306	0.001264	0.001223	0.001183	0.001144	0.001107	0.001070	0.001035	0.001001
-2.9	0.001866	0.001807	0.001750	0.001695	0.001641	0.001589	0.001538	0.001489	0.001441	0.001395
-2.8	0.002555	0.002477	0.002401	0.002327	0.002256	0.002186	0.002118	0.002052	0.001988	0.001926
-2.7	0.003467	0.003364	0.003264	0.003167	0.003072	0.002980	0.002890	0.002803	0.002718	0.002635
-2.6	0.004661	0.004527	0.004396	0.004269	0.004145	0.004025	0.003907	0.003793	0.003681	0.003573
-2.5	0.006210	0.006037	0.005868	0.005703	0.005543	0.005386	0.005234	0.005085	0.004940	0.004799
-2.4	0.008198	0.007976	0.007760	0.007549	0.007344	0.007143	0.006947	0.006756	0.006569	0.006387
-2.3	0.01072	0.01044	0.01017	0.009903	0.009642	0.009387	0.009137	0.008894	0.008656	0.008424
-2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
-2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
-2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
-1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
-1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
-1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
-1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
-1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
-1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
-1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.09853
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

NOTE: Table generated using SAS, a statistical software package. The table entries are values of p, where $p = P(Z \le Z_p)$. For example, $P(Z \le 1.65) = 0.9505$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998									
4.0	1.000									
4.5	1.000									
5.0	1.000									

NOTE: Table generated using SAS, a statistical software package. The table entries are values of p, where $p = P(Z \le Z_p)$. For example, $P(Z \le 1.65) = 0.9505$

Table B-16.
Poisson Probabilities

					μ					
x	**0.005**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0	0.9950	0.9900	0.9802	0.9704	0.9608	0.9512	0.9418	0.9324	0.9231	0.9139
1	0.004975	0.009900	0.01960	0.02911	0.03843	0.04756	0.05651	0.06527	0.07385	0.08225
2	0.00001244	0.00004950	0.0001960	0.0004367	0.0007686	0.001189	0.001695	0.002284	0.002954	0.003701
3	0.00000002073	0.0000001650	0.000001307	0.000004367	0.00001025	0.00001982	0.00003390	0.00005330	0.00007877	0.0001110
x	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.09048	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659	0.3679
2	0.004524	0.01637	0.03334	0.05363	0.07582	0.09879	0.1217	0.1438	0.1647	0.1839
3	0.0001508	0.001092	0.003334	0.007150	0.01264	0.01976	0.02839	0.03834	0.04940	0.06131
4	0.000003770	0.00005458	0.0002500	0.0007150	0.001580	0.002964	0.004968	0.007669	0.01111	0.01533
5	0.00000007540	0.000002183	0.00001500	0.00005720	0.0001580	0.0003556	0.0006955	0.001227	0.002001	0.003066
6	0.000000001257	0.00000007278	0.0000007501	0.000003813	0.00001316	0.00003556	0.00008114	0.0001636	0.0003001	0.0005109
7	0.00000000001795	0.000000002079	0.00000003215	0.0000002179	0.0000009402	0.000003048	0.000008114	0.00001870	0.00003858	0.00007299
x	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**	**2.0**
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353
1	0.3662	0.3614	0.3543	0.3452	0.3347	0.3230	0.3106	0.2975	0.2842	0.2707
2	0.2014	0.2169	0.2303	0.2417	0.2510	0.2584	0.2640	0.2678	0.2700	0.2707
3	0.07384	0.08674	0.09979	0.1128	0.1255	0.1378	0.1496	0.1607	0.1710	0.1804
4	0.02031	0.02602	0.03243	0.03947	0.04707	0.05513	0.06357	0.07230	0.08122	0.09022
5	0.004467	0.006246	0.008432	0.01105	0.01412	0.01764	0.02162	0.02603	0.03086	0.03609
6	0.0008190	0.001249	0.001827	0.002579	0.003530	0.004705	0.006124	0.007809	0.009773	0.01203
7	0.0001287	0.0002141	0.0003393	0.0005158	0.0007564	0.001075	0.001487	0.002008	0.002653	0.003437
8	0.00001770	0.00003212	0.00005514	0.00009026	0.0001418	0.0002151	0.0003161	0.0004518	0.0006300	0.0008593
9	0.000002163	0.000004283	0.000007964	0.00001404	0.00002364	0.00003823	0.00005970	0.00009036	0.0001330	0.0001909
x	**2.1**	**2.2**	**2.3**	**2.4**	**2.5**	**2.6**	**2.7**	**2.8**	**2.9**	**3.0**
0	0.1225	0.1108	0.1003	0.09072	0.08208	0.07427	0.06721	0.06081	0.05502	0.04979
1	0.2572	0.2438	0.2306	0.2177	0.2052	0.1931	0.1815	0.1703	0.1596	0.1494
2	0.2700	0.2681	0.2652	0.2613	0.2565	0.2510	0.2450	0.2384	0.2314	0.2240
3	0.1890	0.1966	0.2033	0.2090	0.2138	0.2176	0.2205	0.2225	0.2237	0.2240
4	0.09923	0.1082	0.1169	0.1254	0.1336	0.1414	0.1488	0.1557	0.1622	0.1680
5	0.04168	0.04759	0.05378	0.06020	0.06680	0.07354	0.08036	0.08721	0.09405	0.1008
6	0.01459	0.01745	0.02061	0.02408	0.02783	0.03187	0.03616	0.04070	0.04546	0.05041
7	0.004376	0.005484	0.006773	0.008255	0.009941	0.01184	0.01395	0.01628	0.01883	0.02160
8	0.001149	0.001508	0.001947	0.002477	0.003106	0.003847	0.004708	0.005698	0.006827	0.008102
9	0.0002680	0.0003686	0.0004976	0.0006604	0.0008629	0.001111	0.001412	0.001773	0.002200	0.002701
10	0.00005629	0.00008110	0.0001145	0.0001585	0.0002157	0.0002889	0.0003813	0.0004964	0.0006379	0.0008102
11	0.00001075	0.00001622	0.00002393	0.00003458	0.00004903	0.00006829	0.00009359	0.0001263	0.0001682	0.0002210
12	0.000001881	0.000002974	0.000004587	0.000006917	0.00001021	0.00001480	0.00002106	0.00002948	0.00004064	0.00005524
x	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**	**3.7**	**3.8**	**3.9**	**4.0**
0	0.04505	0.04076	0.03688	0.03337	0.03020	0.02732	0.02472	0.02237	0.02024	0.01832
1	0.1397	0.1304	0.1217	0.1135	0.1057	0.09837	0.09148	0.08501	0.07894	0.07326
2	0.2165	0.2087	0.2008	0.1929	0.1850	0.1771	0.1692	0.1615	0.1539	0.1465
3	0.2237	0.2226	0.2209	0.2186	0.2158	0.2125	0.2087	0.2046	0.2001	0.1954
4	0.1733	0.1781	0.1823	0.1858	0.1888	0.1912	0.1931	0.1944	0.1951	0.1954
5	0.1075	0.1140	0.1203	0.1264	0.1322	0.1377	0.1429	0.1477	0.1522	0.1563
6	0.05553	0.06079	0.06616	0.07160	0.07710	0.08261	0.08810	0.09355	0.09893	0.1042
7	0.02459	0.02779	0.03119	0.03478	0.03855	0.04248	0.04657	0.05079	0.05512	0.05954
8	0.009529	0.01112	0.01287	0.01478	0.01687	0.01912	0.02154	0.02412	0.02687	0.02977
9	0.003282	0.003952	0.004717	0.005584	0.006559	0.007647	0.008854	0.01019	0.01164	0.01323
10	0.001018	0.001265	0.001557	0.001899	0.002296	0.002753	0.003276	0.003870	0.004541	0.005292
11	0.0002868	0.0003679	0.0004670	0.0005868	0.0007304	0.0009010	0.001102	0.001337	0.001610	0.001925
12	0.00007408	0.00009811	0.0001284	0.0001663	0.0002130	0.0002703	0.0003398	0.0004234	0.0005232	0.0006415
13	0.00001766	0.00002415	0.00003260	0.00004349	0.00005736	0.00007485	0.00009671	0.0001238	0.0001570	0.0001974
14	0.000003911	0.000005520	0.000007684	0.00001056	0.00001434	0.00001925	0.00002556	0.00003359	0.00004373	0.00005640

x	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	0.01657	0.01500	0.01357	0.01228	0.01111	0.01005	0.009095	0.008230	0.007447	0.006738
1	0.06795	0.06298	0.05834	0.05402	0.04999	0.04624	0.04275	0.03950	0.03649	0.03369
2	0.1393	0.1323	0.1254	0.1188	0.1125	0.1063	0.1005	0.09481	0.08940	0.08422
3	0.1904	0.1852	0.1798	0.1743	0.1687	0.1631	0.1574	0.1517	0.1460	0.1404
4	0.1951	0.1944	0.1933	0.1917	0.1898	0.1875	0.1849	0.1820	0.1789	0.1755
5	0.1600	0.1633	0.1662	0.1687	0.1708	0.1725	0.1738	0.1747	0.1753	0.1755
6	0.1093	0.1143	0.1191	0.1237	0.1281	0.1323	0.1362	0.1398	0.1432	0.1462
7	0.06404	0.06859	0.07318	0.07778	0.08236	0.08692	0.09143	0.09586	0.1002	0.1044
8	0.03282	0.03601	0.03933	0.04278	0.04633	0.04998	0.05371	0.05752	0.06138	0.06528
9	0.01495	0.01681	0.01879	0.02091	0.02316	0.02554	0.02805	0.03068	0.03342	0.03627
10	0.006130	0.007058	0.008081	0.009202	0.01042	0.01175	0.01318	0.01472	0.01637	0.01813
11	0.002285	0.002695	0.003159	0.003681	0.004264	0.004914	0.005633	0.006425	0.007294	0.008242
12	0.0007807	0.0009432	0.001132	0.001350	0.001599	0.001884	0.002206	0.002570	0.002978	0.003434
13	0.0002462	0.0003047	0.0003744	0.0004568	0.0005536	0.0006665	0.0007976	0.0009489	0.001123	0.001321
14	0.00007210	0.00009142	0.0001150	0.0001436	0.0001779	0.0002190	0.0002678	0.0003254	0.0003929	0.0004717
15	0.00001971	0.00002560	0.00003297	0.00004211	0.00005338	0.00006716	0.00008390	0.0001041	0.0001284	0.0001572

x	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	0.006097	0.005517	0.004992	0.004517	0.004087	0.003698	0.003346	0.003028	0.002739	0.002479
1	0.03109	0.02869	0.02646	0.02439	0.02248	0.02071	0.01907	0.01756	0.01616	0.01487
2	0.07929	0.07458	0.07011	0.06585	0.06181	0.05798	0.05436	0.05092	0.04768	0.04462
3	0.1348	0.1293	0.1239	0.1185	0.1133	0.1082	0.1033	0.09845	0.09377	0.08924
4	0.1719	0.1681	0.1641	0.1600	0.1558	0.1515	0.1472	0.1428	0.1383	0.1339
5	0.1753	0.1748	0.1740	0.1728	0.1714	0.1697	0.1678	0.1656	0.1632	0.1606
6	0.1490	0.1515	0.1537	0.1555	0.1571	0.1584	0.1594	0.1601	0.1605	0.1606
7	0.1086	0.1125	0.1163	0.1200	0.1234	0.1267	0.1298	0.1326	0.1353	0.1377
8	0.06921	0.07314	0.07708	0.08099	0.08487	0.08870	0.09247	0.09616	0.09976	0.1033
9	0.03922	0.04226	0.04539	0.04859	0.05187	0.05519	0.05856	0.06197	0.06540	0.06884
10	0.02000	0.02198	0.02406	0.02624	0.02853	0.03091	0.03338	0.03594	0.03859	0.04130
11	0.009273	0.01039	0.01159	0.01288	0.01426	0.01573	0.01730	0.01895	0.02070	0.02253
12	0.003941	0.004502	0.005119	0.005797	0.006537	0.007343	0.008216	0.009160	0.01018	0.01126
13	0.001546	0.001801	0.002087	0.002408	0.002766	0.003163	0.003603	0.004087	0.004618	0.005199
14	0.0005632	0.0006688	0.0007901	0.0009288	0.001087	0.001265	0.001467	0.001693	0.001946	0.002228
15	0.0001915	0.0002319	0.0002792	0.0003344	0.0003984	0.0004724	0.0005574	0.0006547	0.0007655	0.0008913
16	0.00006104	0.00007535	0.00009248	0.0001128	0.0001370	0.0001653	0.0001986	0.0002373	0.0002823	0.0003342
17	0.00001831	0.00002305	0.00002883	0.00003585	0.00004431	0.00005446	0.00006658	0.00008097	0.00009797	0.0001180

x	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	0.002243	0.002029	0.001836	0.001662	0.001503	0.001360	0.001231	0.001114	0.001008	0.0009119
1	0.01368	0.01258	0.01157	0.01063	0.009772	0.008978	0.008247	0.007574	0.006954	0.006383
2	0.04173	0.03901	0.03644	0.03403	0.03176	0.02963	0.02763	0.02575	0.02399	0.02234
3	0.08485	0.08061	0.07653	0.07259	0.06881	0.06518	0.06170	0.05837	0.05518	0.05213
4	0.1294	0.1249	0.1205	0.1162	0.1118	0.1076	0.1034	0.09923	0.09518	0.09123
5	0.1579	0.1549	0.1519	0.1487	0.1454	0.1420	0.1385	0.1349	0.1314	0.1277
6	0.1605	0.1601	0.1595	0.1586	0.1575	0.1562	0.1546	0.1529	0.1511	0.1490
7	0.1399	0.1418	0.1435	0.1450	0.1462	0.1472	0.1480	0.1486	0.1489	0.1490
8	0.1066	0.1099	0.1130	0.1160	0.1188	0.1215	0.1240	0.1263	0.1284	0.1304
9	0.07228	0.07571	0.07911	0.08248	0.08581	0.08908	0.09229	0.09541	0.09846	0.1014
10	0.04409	0.04694	0.04984	0.05279	0.05578	0.05879	0.06183	0.06488	0.06794	0.07098
11	0.02445	0.02646	0.02855	0.03071	0.03296	0.03528	0.03766	0.04011	0.04261	0.04517
12	0.01243	0.01367	0.01499	0.01638	0.01785	0.01940	0.02103	0.02273	0.02450	0.02635
13	0.005832	0.006519	0.007263	0.008064	0.008926	0.009850	0.01084	0.01189	0.01301	0.01419
14	0.002541	0.002887	0.003268	0.003687	0.004144	0.004644	0.005186	0.005774	0.006410	0.007094
15	0.001033	0.001193	0.001373	0.001573	0.001796	0.002043	0.002317	0.002618	0.002949	0.003311
16	0.0003940	0.0004624	0.0005405	0.0006292	0.0007296	0.0008428	0.0009701	0.001113	0.001272	0.001448
17	0.0001414	0.0001686	0.0002003	0.0002369	0.0002790	0.0003272	0.0003823	0.0004450	0.0005161	0.0005964
18	0.00004791	0.00005809	0.00007010	0.00008422	0.0001007	0.0001200	0.0001423	0.0001681	0.0001978	0.0002319
19	0.00001538	0.00001895	0.00002324	0.00002837	0.00003446	0.00004168	0.00005018	0.00006017	0.00007185	0.00008545

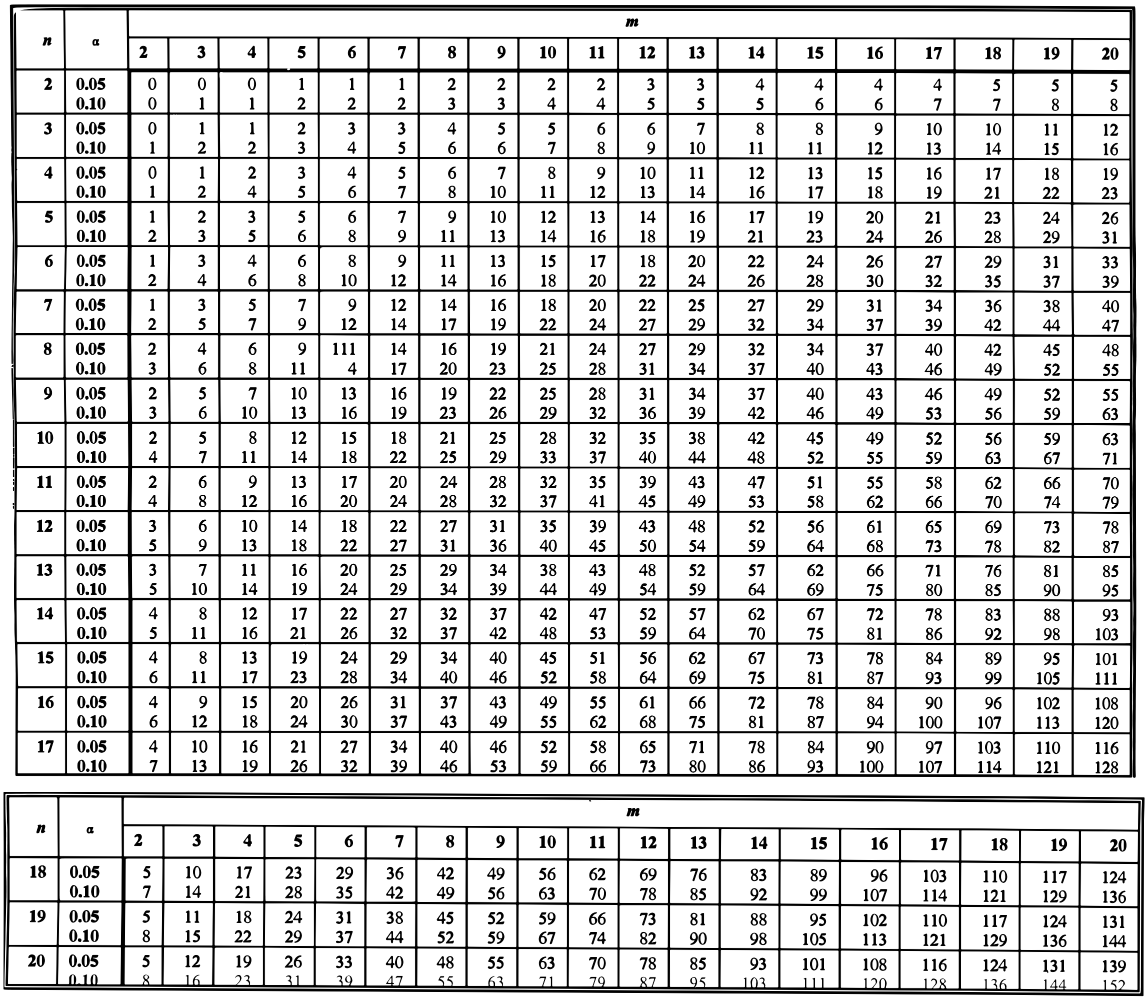
x	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	0.0008251	0.0007466	0.0006755	0.0006113	0.0005531	0.0005005	0.0004528	0.0004097	0.0003707	0.0003355
1	0.005858	0.005375	0.004931	0.004523	0.004148	0.003803	0.003487	0.003196	0.002929	0.002684
2	0.02080	0.01935	0.01800	0.01674	0.01556	0.01445	0.01342	0.01246	0.01157	0.01073
3	0.04922	0.04644	0.04380	0.04128	0.03889	0.03661	0.03446	0.03241	0.03047	0.02863
4	0.08736	0.08360	0.07993	0.07637	0.07292	0.06957	0.06633	0.06319	0.06017	0.05725
5	0.1241	0.1204	0.1167	0.1130	0.1094	0.1057	0.1021	0.09858	0.09507	0.09160
6	0.1468	0.1445	0.1420	0.1394	0.1367	0.1339	0.1311	0.1282	0.1252	0.1221
7	0.1489	0.1486	0.1481	0.1474	0.1465	0.1454	0.1442	0.1428	0.1413	0.1396
8	0.1321	0.1337	0.1351	0.1363	0.1373	0.1381	0.1388	0.1392	0.1395	0.1396
9	0.1042	0.1070	0.1096	0.1121	0.1144	0.1167	0.1187	0.1207	0.1224	0.1241
10	0.07402	0.07703	0.08000	0.08294	0.08583	0.08866	0.09143	0.09412	0.09673	0.09926
11	0.04777	0.05042	0.05309	0.05580	0.05852	0.06126	0.06400	0.06674	0.06947	0.07219
12	0.02827	0.03025	0.03230	0.03441	0.03658	0.03880	0.04107	0.04338	0.04574	0.04813
13	0.01544	0.01675	0.01814	0.01959	0.02110	0.02268	0.02432	0.02603	0.02779	0.02962
14	0.007829	0.008616	0.009457	0.01035	0.01130	0.01231	0.01338	0.01450	0.01568	0.01692
15	0.003706	0.004136	0.004602	0.005107	0.005652	0.006238	0.006867	0.007541	0.008260	0.009026
16	0.001644	0.001861	0.002100	0.002362	0.002649	0.002963	0.003305	0.003676	0.004078	0.004513
17	0.0006868	0.0007882	0.0009017	0.001028	0.001169	0.001325	0.001497	0.001687	0.001895	0.002124
18	0.0002709	0.0003153	0.0003657	0.0004227	0.0004870	0.0005593	0.0006404	0.0007309	0.0008318	0.0009439
19	0.0001012	0.0001195	0.0001405	0.0001646	0.0001922	0.0002237	0.0002595	0.0003001	0.0003459	0.0003974
20	0.00003594	0.00004301	0.00005128	0.00006092	0.00007209	0.00008502	0.00009991	0.0001170	0.0001366	0.0001590
21	0.00001215	0.00001475	0.00001783	0.00002147	0.00002575	0.00003077	0.00003663	0.00004347	0.00005139	0.00006056

x	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	0.0003035	0.0002747	0.0002485	0.0002249	0.0002035	0.0001841	0.0001666	0.0001507	0.0001364	0.0001234
1	0.002459	0.002252	0.002063	0.001889	0.001729	0.001583	0.001449	0.001326	0.001214	0.001111
2	0.009958	0.009234	0.008560	0.007933	0.007350	0.006808	0.006304	0.005836	0.005402	0.004998
3	0.02689	0.02524	0.02368	0.02221	0.02083	0.01952	0.01828	0.01712	0.01602	0.01499
4	0.05444	0.05174	0.04914	0.04665	0.04425	0.04196	0.03977	0.03766	0.03566	0.03374
5	0.08820	0.08485	0.08158	0.07837	0.07523	0.07217	0.06919	0.06629	0.06347	0.06073
6	0.1191	0.1160	0.1128	0.1097	0.1066	0.1034	0.1003	0.09722	0.09414	0.09109
7	0.1378	0.1358	0.1338	0.1317	0.1294	0.1271	0.1247	0.1222	0.1197	0.1171
8	0.1395	0.1392	0.1388	0.1382	0.1375	0.1366	0.1356	0.1344	0.1332	0.1318
9	0.1256	0.1269	0.1280	0.1290	0.1299	0.1306	0.1311	0.1315	0.1317	0.1318
10	0.1017	0.1040	0.1063	0.1084	0.1104	0.1123	0.1140	0.1157	0.1172	0.1186
11	0.07488	0.07755	0.08018	0.08276	0.08530	0.08778	0.09020	0.09255	0.09482	0.09702
12	0.05055	0.05299	0.05546	0.05793	0.06042	0.06291	0.06539	0.06787	0.07033	0.07277
13	0.03149	0.03343	0.03541	0.03743	0.03951	0.04162	0.04376	0.04594	0.04815	0.05038
14	0.01822	0.01958	0.02099	0.02246	0.02399	0.02556	0.02720	0.02888	0.03061	0.03238
15	0.009840	0.01070	0.01162	0.01258	0.01359	0.01466	0.01577	0.01694	0.01816	0.01943
16	0.004981	0.005485	0.006025	0.006604	0.007221	0.007878	0.008577	0.009318	0.01010	0.01093
17	0.002374	0.002646	0.002942	0.003263	0.003610	0.003985	0.004389	0.004823	0.005289	0.005786
18	0.001068	0.001205	0.001357	0.001523	0.001705	0.001904	0.002122	0.002358	0.002615	0.002893
19	0.0004553	0.0005202	0.0005926	0.0006732	0.0007627	0.0008619	0.0009714	0.001092	0.001225	0.001370
20	0.0001844	0.0002133	0.0002459	0.0002827	0.0003242	0.0003706	0.0004226	0.0004805	0.0005451	0.0006167
21	0.00007113	0.00008328	0.00009720	0.0001131	0.0001312	0.0001518	0.0001751	0.0002014	0.0002310	0.0002643
22	0.00002619	0.00003104	0.00003667	0.00004318	0.00005069	0.00005933	0.00006923	0.00008055	0.00009345	0.0001081

x	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10.0
0	0.0001117	0.0001010	0.00009142	0.00008272	0.00007485	0.00006773	0.00006128	0.00005545	0.00005017	0.00004540
1	0.001016	0.0009296	0.0008502	0.0007776	0.0007111	0.0006502	0.0005944	0.0005434	0.0004967	0.0004540
2	0.004624	0.004276	0.003954	0.003655	0.003378	0.003121	0.002883	0.002663	0.002459	0.002270
3	0.01402	0.01311	0.01226	0.01145	0.01070	0.009987	0.009322	0.008698	0.008114	0.007567
4	0.03191	0.03016	0.02850	0.02691	0.02540	0.02397	0.02261	0.02131	0.02008	0.01892
5	0.05807	0.05549	0.05300	0.05059	0.04827	0.04602	0.04386	0.04177	0.03976	0.03783
6	0.08807	0.08509	0.08215	0.07926	0.07642	0.07363	0.07090	0.06822	0.06561	0.06306
7	0.1145	0.1118	0.1091	0.1064	0.1037	0.1010	0.09825	0.09551	0.09279	0.09008
8	0.1302	0.1286	0.1269	0.1251	0.1232	0.1212	0.1191	0.1170	0.1148	0.1126
9	0.1317	0.1315	0.1311	0.1306	0.1300	0.1293	0.1284	0.1274	0.1263	0.1251
10	0.1198	0.1210	0.1219	0.1228	0.1235	0.1241	0.1245	0.1249	0.1250	0.1251
11	0.09913	0.1012	0.1031	0.1049	0.1067	0.1083	0.1098	0.1112	0.1125	0.1137
12	0.07518	0.07755	0.07990	0.08219	0.08444	0.08663	0.08877	0.09084	0.09285	0.09478
13	0.05262	0.05488	0.05716	0.05943	0.06171	0.06398	0.06624	0.06848	0.07071	0.07291
14	0.03421	0.03607	0.03797	0.03990	0.04187	0.04387	0.04589	0.04794	0.05000	0.05208
15	0.02075	0.02212	0.02354	0.02501	0.02652	0.02808	0.02968	0.03132	0.03300	0.03472
16	0.01180	0.01272	0.01368	0.01469	0.01575	0.01685	0.01799	0.01918	0.02042	0.02170
17	0.006318	0.006884	0.007485	0.008123	0.008799	0.009513	0.01027	0.01106	0.01189	0.01276
18	0.003194	0.003518	0.003867	0.004242	0.004644	0.005074	0.005532	0.006021	0.006540	0.007091
19	0.001530	0.001704	0.001893	0.002099	0.002322	0.002563	0.002824	0.003105	0.003408	0.003732
20	0.0006960	0.0007837	0.0008802	0.0009864	0.001103	0.001230	0.001370	0.001522	0.001687	0.001866
21	0.0003016	0.0003433	0.0003898	0.0004415	0.0004989	0.0005625	0.0006327	0.0007101	0.0007952	0.0008886
22	0.0001248	0.0001436	0.0001648	0.0001887	0.0002155	0.0002455	0.0002790	0.0003163	0.0003578	0.0004039
23	0.00004936	0.00005743	0.00006663	0.00007710	0.00008899	0.0001025	0.0001177	0.0001348	0.0001540	0.0001756
24	0.00001872	0.00002201	0.00002582	0.00003020	0.00003523	0.00004098	0.00004755	0.00005503	0.00006354	0.00007317

Source (: Kvanli et al. (1996).

Table B-17.
Critical Values for the Rank-Sum Test

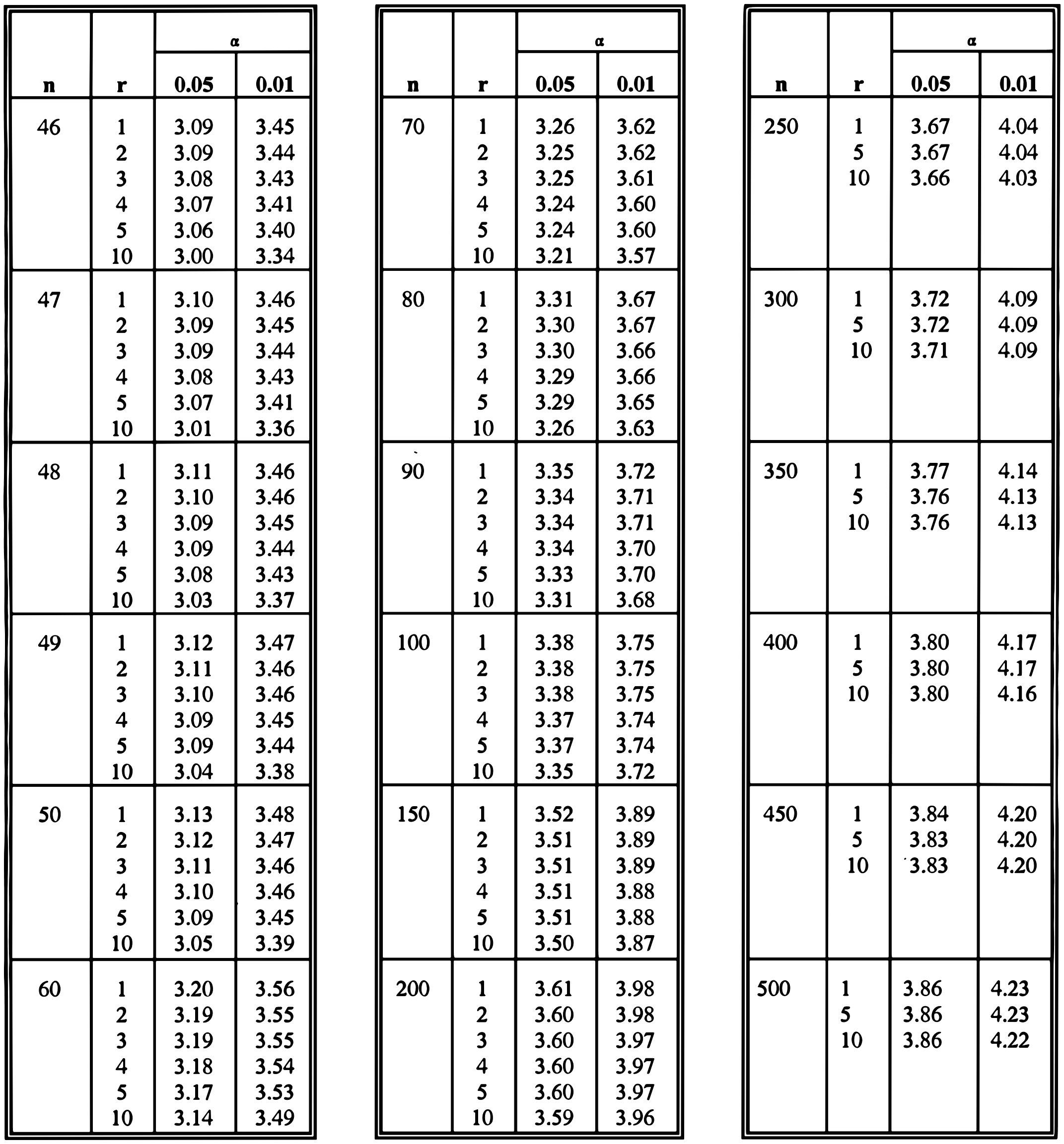


n	α	*m*																		
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	0.05	0	0	0	1	1	1	2	2	2	2	3	3	4	4	4	4	5	5	5
	0.10	0	1	1	2	2	2	3	3	4	4	5	5	5	6	6	7	7	8	8
3	0.05	0	1	1	2	3	3	4	5	5	6	6	7	8	8	9	10	10	11	12
	0.10	1	2	2	3	4	5	6	6	7	8	9	10	11	11	12	13	14	15	16
4	0.05	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19
	0.10	1	2	4	5	6	7	8	10	11	12	13	14	16	17	18	19	21	22	23
5	0.05	1	2	3	5	6	7	9	10	12	13	14	16	17	19	20	21	23	24	26
	0.10	2	3	5	6	8	9	11	13	14	16	18	19	21	23	24	26	28	29	31
6	0.05	1	3	4	6	8	9	11	13	15	17	18	20	22	24	26	27	29	31	33
	0.10	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	35	37	39
7	0.05	1	3	5	7	9	12	14	16	18	20	22	25	27	29	31	34	36	38	40
	0.10	2	5	7	9	12	14	17	19	22	24	27	29	32	34	37	39	42	44	47
8	0.05	2	4	6	9	111	14	16	19	21	24	27	29	32	34	37	40	42	45	48
	0.10	3	6	8	11	4	17	20	23	25	28	31	34	37	40	43	46	49	52	55
9	0.05	2	5	7	10	13	16	19	22	25	28	31	34	37	40	43	46	49	52	55
	0.10	3	6	10	13	16	19	23	26	29	32	36	39	42	46	49	53	56	59	63
10	0.05	2	5	8	12	15	18	21	25	28	32	35	38	42	45	49	52	56	59	63
	0.10	4	7	11	14	18	22	25	29	33	37	40	44	48	52	55	59	63	67	71
11	0.05	2	6	9	13	17	20	24	28	32	35	39	43	47	51	55	58	62	66	70
	0.10	4	8	12	16	20	24	28	32	37	41	45	49	53	58	62	66	70	74	79
12	0.05	3	6	10	14	18	22	27	31	35	39	43	48	52	56	61	65	69	73	78
	0.10	5	9	13	18	22	27	31	36	40	45	50	54	59	64	68	73	78	82	87
13	0.05	3	7	11	16	20	25	29	34	38	43	48	52	57	62	66	71	76	81	85
	0.10	5	10	14	19	24	29	34	39	44	49	54	59	64	69	75	80	85	90	95
14	0.05	4	8	12	17	22	27	32	37	42	47	52	57	62	67	72	78	83	88	93
	0.10	5	11	16	21	26	32	37	42	48	53	59	64	70	75	81	86	92	98	103
15	0.05	4	8	13	19	24	29	34	40	45	51	56	62	67	73	78	84	89	95	101
	0.10	6	11	17	23	28	34	40	46	52	58	64	69	75	81	87	93	99	105	111
16	0.05	4	9	15	20	26	31	37	43	49	55	61	66	72	78	84	90	96	102	108
	0.10	6	12	18	24	30	37	43	49	55	62	68	75	81	87	94	100	107	113	120
17	0.05	4	10	16	21	27	34	40	46	52	58	65	71	78	84	90	97	103	110	116
	0.10	7	13	19	26	32	39	46	53	59	66	73	80	86	93	100	107	114	121	128
18	0.05	5	10	17	23	29	36	42	49	56	62	69	76	83	89	96	103	110	117	124
	0.10	7	14	21	28	35	42	49	56	63	70	78	85	92	99	107	114	121	129	136
19	0.05	5	11	18	24	31	38	45	52	59	66	73	81	88	95	102	110	117	124	131
	0.10	8	15	22	29	37	44	52	59	67	74	82	90	98	105	113	121	129	136	144
20	0.05	5	12	19	26	33	40	48	55	63	70	78	85	93	101	108	116	124	131	139
	0.10	8	16	23	31	39	47	55	63	71	79	87	95	103	111	120	128	136	144	152

Source: EPA/600/R-96/084.

Table B-18.
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
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
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



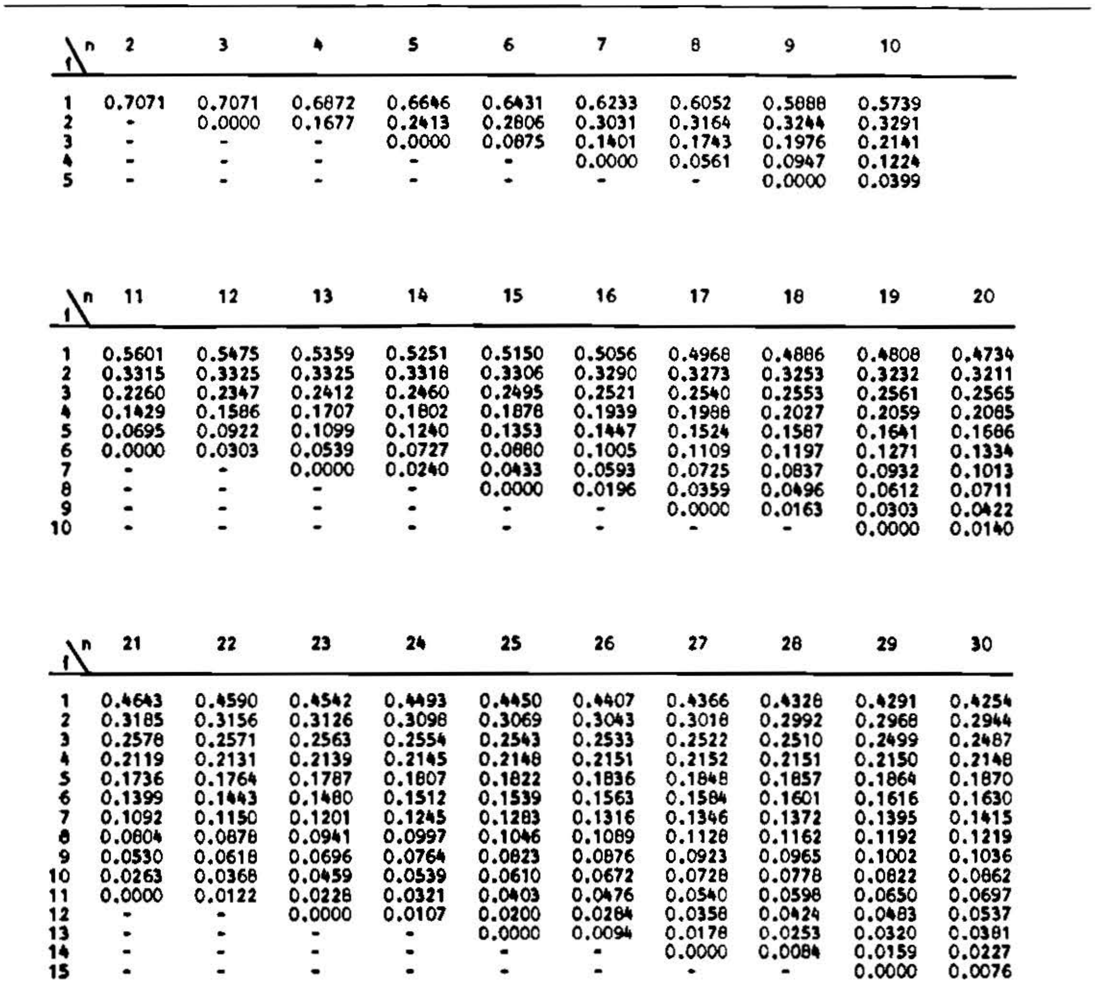
n	r	α	
		0.05	0.01
46	1	3.09	3.45
	2	3.09	3.44
	3	3.08	3.43
	4	3.07	3.41
	5	3.06	3.40
	10	3.00	3.34
47	1	3.10	3.46
	2	3.09	3.45
	3	3.09	3.44
	4	3.08	3.43
	5	3.07	3.41
	10	3.01	3.36
48	1	3.11	3.46
	2	3.10	3.46
	3	3.09	3.45
	4	3.09	3.44
	5	3.08	3.43
	10	3.03	3.37
49	1	3.12	3.47
	2	3.11	3.46
	3	3.10	3.46
	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

n	r	α	
		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

n	r	α	
		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

Source: EPA/600/R-96/084.

Table B-19.
Coefficients for the Shapiro-Wilk W Test for Normality



i \ n	2	3	4	5	6	7	8	9	10
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739
2	-	0.0000	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291
3	-	-	-	0.0000	0.0875	0.1401	0.1743	0.1976	0.2141
4	-	-	-	-	-	0.0000	0.0561	0.0947	0.1224
5	-	-	-	-	-	-	-	0.0000	0.0399

i \ n	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7	-	-	0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8	-	-	-	-	0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9	-	-	-	-	-	-	0.0000	0.0163	0.0303	0.0422
10	-	-	-	-	-	-	-	-	0.0000	0.0140

i \ n	21	22	23	24	25	26	27	28	29	30
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254
2	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043	0.3018	0.2992	0.2968	0.2944
3	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533	0.2522	0.2510	0.2499	0.2487
4	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151	0.2152	0.2151	0.2150	0.2148
5	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836	0.1848	0.1857	0.1864	0.1870
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630
7	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316	0.1346	0.1372	0.1395	0.1415
8	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089	0.1128	0.1162	0.1192	0.1219
9	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876	0.0923	0.0965	0.1002	0.1036
10	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672	0.0728	0.0778	0.0822	0.0862
11	0.0000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697
12	-	-	0.0000	0.0107	0.0200	0.0284	0.0358	0.0424	0.0483	0.0537
13	-	-	-	-	0.0000	0.0094	0.0178	0.0253	0.0320	0.0381
14	-	-	-	-	-	-	0.0000	0.0084	0.0159	0.0227
15	-	-	-	-	-	-	-	-	0.0000	0.0076

Source: From Shapiro and Wilk, 1965. Used by permission.
This table is used in Section 12.3.1.

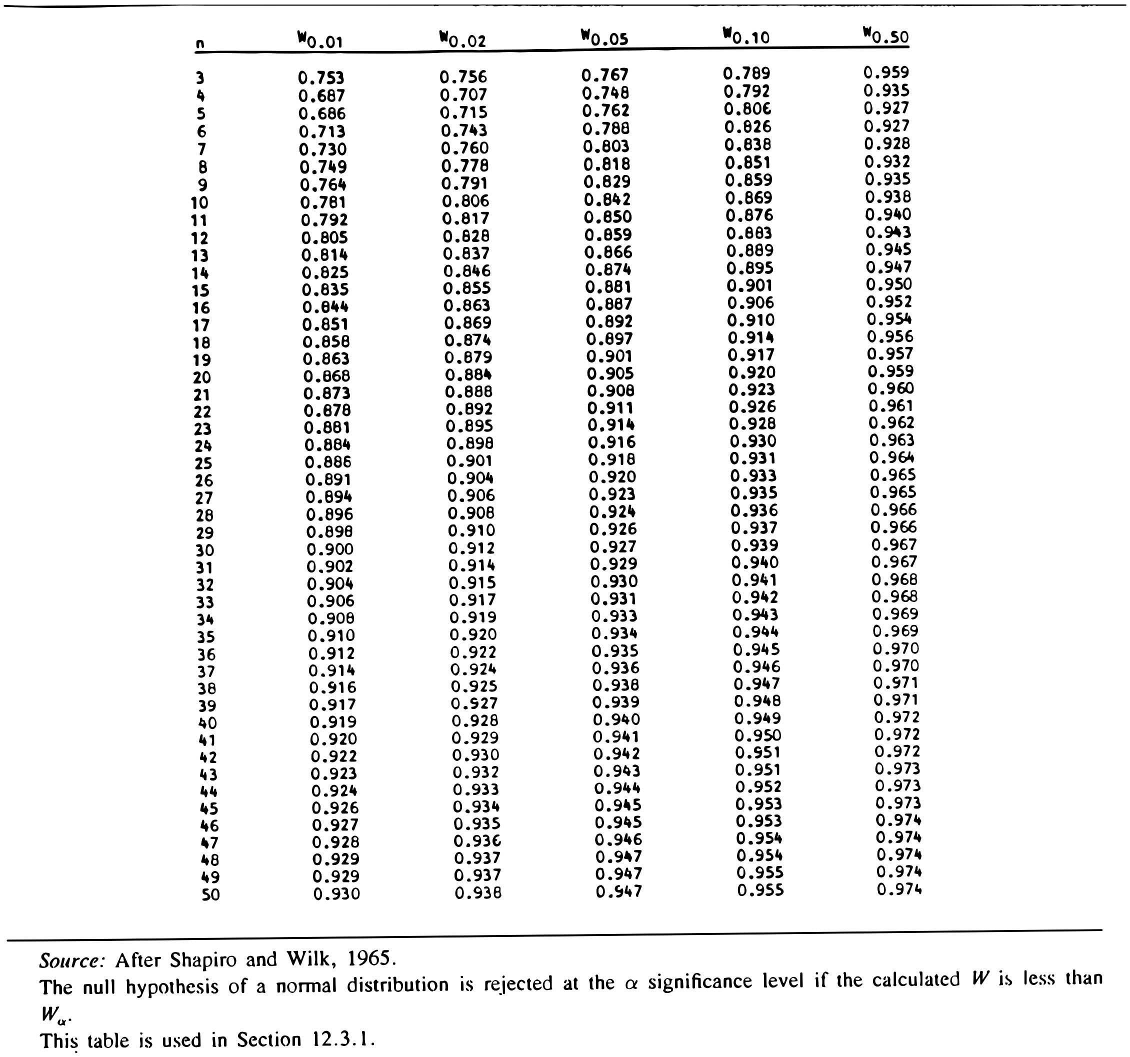
i \ n	31	32	33	34	35	36	37	38	39	40
1	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015	0.3989	0.3964
2	0.2921	0.2898	0.2876	0.2854	0.2834	0.2813	0.2794	0.2774	0.2755	0.2737
3	0.2475	0.2462	0.2451	0.2439	0.2427	0.2415	0.2403	0.2391	0.2380	0.2368
4	0.2145	0.2141	0.2137	0.2132	0.2127	0.2121	0.2116	0.2110	0.2104	0.2098
5	0.1874	0.1878	0.1880	0.1882	0.1883	0.1883	0.1883	0.1881	0.1880	0.1878
6	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686	0.1689	0.1691
7	0.1433	0.1449	0.1463	0.1475	0.1487	0.1496	0.1505	0.1513	0.1520	0.1526
8	0.1243	0.1265	0.1284	0.1301	0.1317	0.1331	0.1344	0.1356	0.1366	0.1376
9	0.1066	0.1093	0.1116	0.1140	0.1160	0.1179	0.1196	0.1211	0.1225	0.1237
10	0.0899	0.0931	0.0961	0.0988	0.1013	0.1036	0.1056	0.1075	0.1092	0.1108
11	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947	0.0967	0.0986
12	0.0585	0.0629	0.0669	0.0706	0.0739	0.0770	0.0798	0.0824	0.0848	0.0870
13	0.0435	0.0485	0.0530	0.0572	0.0610	0.0645	0.0677	0.0706	0.0733	0.0759
14	0.0289	0.0344	0.0395	0.0441	0.0484	0.0523	0.0559	0.0592	0.0622	0.0651
15	0.0144	0.0206	0.0262	0.0314	0.0361	0.0404	0.0444	0.0481	0.0515	0.0546
16	0.0000	0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372	0.0409	0.0444
17	-	-	0.0000	0.0062	0.0119	0.0172	0.0220	0.0264	0.0305	0.0343
18	-	-	-	-	0.0000	0.0057	0.0110	0.0158	0.0203	0.0244
19	-	-	-	-	-	-	0.0000	0.0053	0.0101	0.0146
20	-	-	-	-	-	-	-	-	0.0000	0.0049

i \ n	41	42	43	44	45	46	47	48	49	50
1	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	0.2719	0.2701	0.2684	0.2667	0.2651	0.2635	0.2620	0.2604	0.2589	0.2574
3	0.2357	0.2345	0.2334	0.2323	0.2313	0.2302	0.2291	0.2281	0.2271	0.2260
4	0.2091	0.2085	0.2078	0.2072	0.2065	0.2058	0.2052	0.2045	0.2038	0.2032
5	0.1876	0.1874	0.1871	0.1868	0.1865	0.1862	0.1859	0.1855	0.1851	0.1847
6	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	0.1531	0.1535	0.1539	0.1542	0.1545	0.1548	0.1550	0.1551	0.1553	0.1554
8	0.1384	0.1392	0.1398	0.1405	0.1410	0.1415	0.1420	0.1423	0.1427	0.1430
9	0.1249	0.1259	0.1269	0.1278	0.1286	0.1293	0.1300	0.1306	0.1312	0.1317
10	0.1123	0.1136	0.1149	0.1160	0.1170	0.1180	0.1189	0.1197	0.1205	0.1212
11	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	0.0891	0.0909	0.0927	0.0943	0.0959	0.0972	0.0986	0.0998	0.1010	0.1020
13	0.0782	0.0804	0.0824	0.0842	0.0860	0.0876	0.0892	0.0906	0.0919	0.0932
14	0.0677	0.0701	0.0724	0.0745	0.0765	0.0783	0.0801	0.0817	0.0832	0.0846
15	0.0575	0.0602	0.0628	0.0651	0.0673	0.0694	0.0713	0.0731	0.0748	0.0764
16	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	0.0379	0.0411	0.0442	0.0471	0.0497	0.0522	0.0546	0.0568	0.0588	0.0608
18	0.0283	0.0318	0.0352	0.0383	0.0412	0.0439	0.0465	0.0489	0.0511	0.0532
19	0.0188	0.0227	0.0263	0.0296	0.0328	0.0357	0.0385	0.0411	0.0436	0.0459
20	0.0094	0.0136	0.0175	0.0211	0.0245	0.0277	0.0307	0.0335	0.0361	0.0386
21	0.0000	0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22	-	-	0.0000	0.0042	0.0081	0.0118	0.0153	0.0185	0.0215	0.0244
23	-	-	-	-	0.0000	0.0039	0.0076	0.0111	0.0143	0.0174
24	-	-	-	-	-	-	0.0000	0.0037	0.0071	0.0104
25	-	-	-	-	-	-	-	-	0.0000	0.0035

Source: Gilbert (1987).

Note: The coefficients listed in the table are denoted as $a_{(n-i+1)}$ in Appendix F. For the value of n listed on the top of each column, the rows list the values of $a_{(n-i+1)}$, where $i = 1, \ldots, k$ and k is the largest integer less than or equal to $n/2$.

Table B-20.
Quantiles W_α of the Shapiro-Wilk W Test for Normality



n	$W_{0.01}$	$W_{0.02}$	$W_{0.05}$	$W_{0.10}$	$W_{0.50}$
3	0.753	0.756	0.767	0.789	0.959
4	0.687	0.707	0.748	0.792	0.935
5	0.686	0.715	0.762	0.806	0.927
6	0.713	0.743	0.788	0.826	0.927
7	0.730	0.760	0.803	0.838	0.928
8	0.749	0.778	0.818	0.851	0.932
9	0.764	0.791	0.829	0.859	0.935
10	0.781	0.806	0.842	0.869	0.938
11	0.792	0.817	0.850	0.876	0.940
12	0.805	0.828	0.859	0.883	0.943
13	0.814	0.837	0.866	0.889	0.945
14	0.825	0.846	0.874	0.895	0.947
15	0.835	0.855	0.881	0.901	0.950
16	0.844	0.863	0.887	0.906	0.952
17	0.851	0.869	0.892	0.910	0.954
18	0.858	0.874	0.897	0.914	0.956
19	0.863	0.879	0.901	0.917	0.957
20	0.868	0.884	0.905	0.920	0.959
21	0.873	0.888	0.908	0.923	0.960
22	0.878	0.892	0.911	0.926	0.961
23	0.881	0.895	0.914	0.928	0.962
24	0.884	0.898	0.916	0.930	0.963
25	0.886	0.901	0.918	0.931	0.964
26	0.891	0.904	0.920	0.933	0.965
27	0.894	0.906	0.923	0.935	0.965
28	0.896	0.908	0.924	0.936	0.966
29	0.898	0.910	0.926	0.937	0.966
30	0.900	0.912	0.927	0.939	0.967
31	0.902	0.914	0.929	0.940	0.967
32	0.904	0.915	0.930	0.941	0.968
33	0.906	0.917	0.931	0.942	0.968
34	0.908	0.919	0.933	0.943	0.969
35	0.910	0.920	0.934	0.944	0.969
36	0.912	0.922	0.935	0.945	0.970
37	0.914	0.924	0.936	0.946	0.970
38	0.916	0.925	0.938	0.947	0.971
39	0.917	0.927	0.939	0.948	0.971
40	0.919	0.928	0.940	0.949	0.972
41	0.920	0.929	0.941	0.950	0.972
42	0.922	0.930	0.942	0.951	0.972
43	0.923	0.932	0.943	0.951	0.973
44	0.924	0.933	0.944	0.952	0.973
45	0.926	0.934	0.945	0.953	0.973
46	0.927	0.935	0.945	0.953	0.974
47	0.928	0.936	0.946	0.954	0.974
48	0.929	0.937	0.947	0.954	0.974
49	0.929	0.937	0.947	0.955	0.974
50	0.930	0.938	0.947	0.955	0.974

Source: After Shapiro and Wilk, 1965.
The null hypothesis of a normal distribution is rejected at the α significance level if the calculated W is less than W_α.
This table is used in Section 12.3.1.

Source: Gilbert (1987).

Note: The assumption of normality is rejected at the $(1 - \alpha)100\%$ level of confidence when the calculated value of $W < W_\alpha$, where $P(W \leq W_\alpha) = \alpha$.

Table B-21.
Critical Values for the Studentized Range Test

	Level of Significance, α					
	0.01		0.05		0.1	
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

Source: EPA/600/R-96/084.

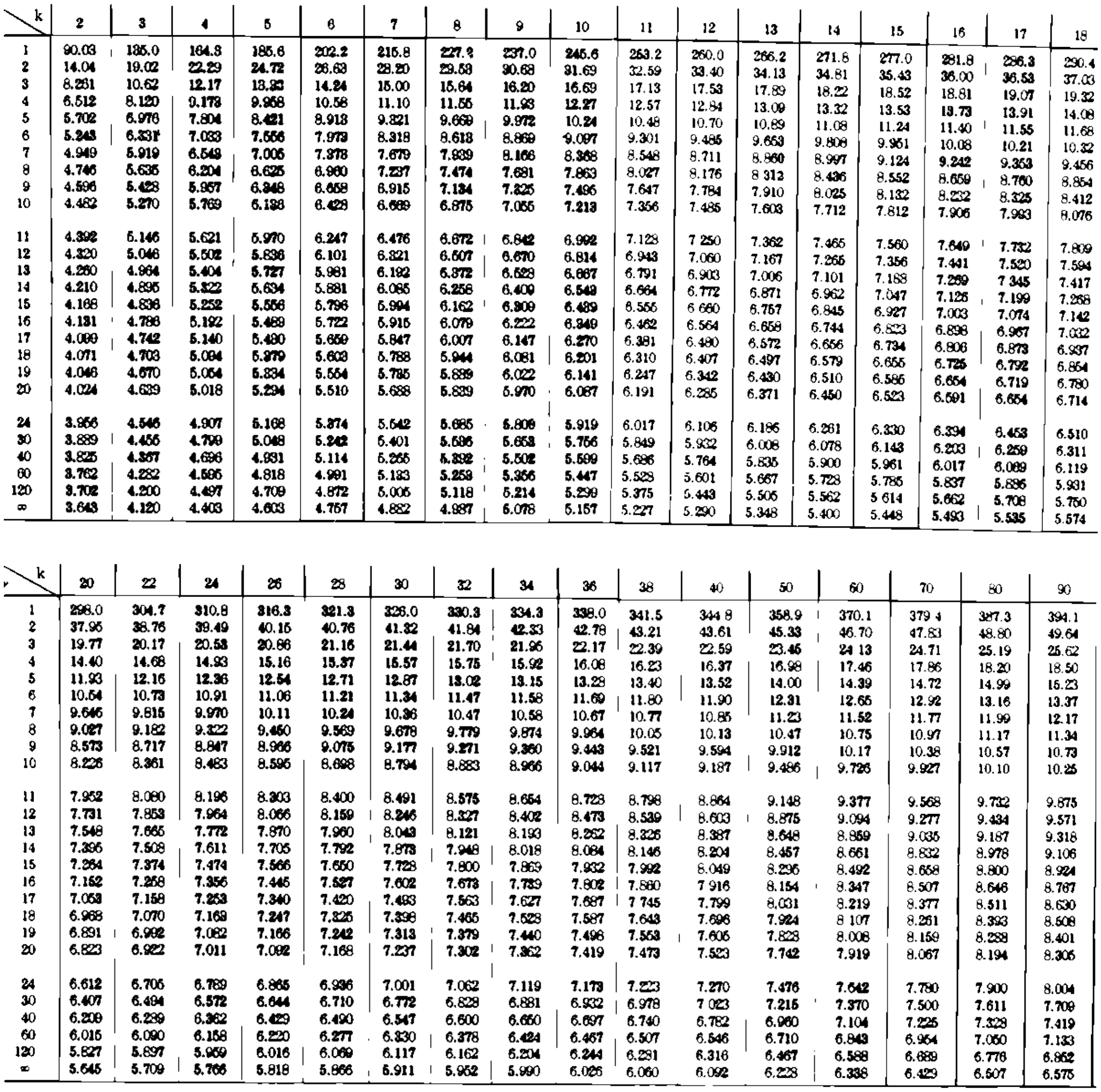
Table B-22.
Percentage Points of the Studentized Range

$\alpha = 0.05$

k	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07	50.59	51.96	53.20	54.33	55.36	56.32	57.22	58.04
2	6.085	8.331	9.798	10.88	11.74	12.44	13.03	13.54	13.99	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.37
3	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462	9.717	9.946	10.15	10.35	10.53	10.69	10.84	10.98
4	3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.602	7.826	8.027	8.208	8.373	8.525	8.664	8.794	8.914	9.028
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.802	6.995	7.168	7.324	7.466	7.596	7.717	7.828	7.932	8.030
6	3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493	6.649	6.789	6.917	7.034	7.143	7.244	7.338	7.426
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.998	6.158	6.302	6.431	6.550	6.658	6.759	6.852	6.939	7.020
8	3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918	6.054	6.175	6.287	6.389	6.483	6.571	6.653	6.729
9	3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739	5.867	5.983	6.089	6.186	6.276	6.359	6.437	6.510
10	3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599	5.722	5.833	5.935	6.028	6.114	6.194	6.269	6.339
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487	5.605	5.713	5.811	5.901	5.984	6.062	6.134	6.202
12	3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395	5.511	5.615	5.710	5.798	5.878	5.953	6.023	6.089
13	3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318	5.431	5.533	5.625	5.711	5.789	5.862	5.931	5.995
14	3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254	5.364	5.463	5.554	5.637	5.714	5.785	5.852	5.915
15	3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198	5.306	5.404	5.493	5.574	5.649	5.720	5.785	5.846
16	2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150	5.256	5.352	5.439	5.520	5.593	5.662	5.727	5.786
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108	5.212	5.307	5.392	5.471	5.544	5.612	5.675	5.734
18	2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071	5.174	5.267	5.352	5.429	5.501	5.568	5.630	5.688
19	2.960	3.593	3.977	4.253	4.469	4.466	4.794	4.924	5.038	5.140	5.231	5.315	5.391	5.462	5.528	5.589	5.647
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008	5.108	5.199	5.282	5.357	5.427	5.493	5.553	5.610
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915	5.012	5.099	5.179	5.251	5.319	5.381	5.439	5.494
30	2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824	4.917	5.001	5.077	5.147	5.211	5.271	5.327	5.379
40	2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735	4.824	4.904	4.977	5.044	5.106	5.163	5.216	5.266
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646	4.732	4.808	4.878	4.942	5.001	5.056	5.107	5.154
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560	4.641	4.714	4.781	4.842	4.898	4.950	4.998	5.044
∞	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474	4.552	4.622	4.685	4.743	4.796	4.845	4.891	4.934

ν k	20	22	24	26	28	30	32	34	36	38	40	50	60	70	80	90
1	59.56	60.91	62.12	63.22	64.23	65.15	66.01	66.81	67.56	68.26	68.92	71 73	73.97	75.82	77.40	78.77
2	16.77	17.13	17.45	17.75	18.03	18.27	18.50	18.72	18.92	19.11	19.28	20.05	20.66	21.16	21.59	21.96
3	11.24	11.47	11.68	11.87	12.05	12.21	12.36	12.50	12.63	12.75	12.87	13.36	13.76	14.08	14.36	14.61
4	9.233	9.418	9.584	9.736	9.875	10.00	10.12	10.23	10.34	10.44	10.53	10.93	11.24	11.51	11.73	11.92
5	8.208	8.368	8.512	8.643	8.764	8.875	8.979	9.075	9.165	9.250	9.330	9.674	9.949	10.18	10.38	10.54
6	7.587	7.730	7.861	7.979	8.088	8.189	8.283	8.370	8.452	8.529	8.601	8.913	9.163	9.370	9.548	9.702
7	7.170	7.303	7.423	7.533	7.634	7.728	7.814	7.895	7.972	8.043	8.110	8.400	8.632	8.824	8.989	9.133
8	6.870	6.995	7.109	7.212	7.307	7.395	7.477	7.554	7.625	7.693	7.756	8.029	8.248	8.430	8.586	8.722
9	6.644	6.763	6.871	6.970	7.061	7.145	7.222	7.295	7.363	7.428	7.488	7.749	7.958	8.132	8.281	8.410
10	6.467	6.582	6.686	6.781	6.868	6.948	7.023	7.093	7.159	7.220	7.279	7.529	7.730	7.897	8.041	8.166
11	6.326	6.436	6.536	6.628	6.712	6.790	6.863	6.930	6.994	7.053	7.110	7.352	7.546	7.708	7.847	7.968
12	6.209	6.317	6.414	6.503	6.585	6.660	6.731	6.796	6.858	6.916	6.970	7.205	7.394	7.552	7.687	7.804
13	6.112	6.217	6.312	6.398	6.478	6.551	6.620	6.684	6.744	6.800	6.854	7.083	7.267	7.421	7.552	7.667
14	6.029	6.132	6.224	6.309	6.387	6.459	6.526	6.588	6.647	6.702	6.754	6.979	7.159	7.309	7.438	7.550
15	5.958	6.059	6.149	6.233	6.309	6.379	6.445	6.506	6.564	6.618	6.669	6.888	7.065	7.212	7.339	7.449
16	5.897	5.995	6.084	6.166	6.241	6.310	6.374	6.434	6.491	6.544	6.594	6.810	6.984	7.128	7.252	7.360
17	5.842	5.940	6.027	6.107	6.181	6.249	6.312	6.372	6.427	6.479	6.529	6.741	6.912	7.054	7.176	7.283
18	5.794	5.890	5.977	6.055	6.127	6.195	6.258	6.316	6.371	6.422	6.471	6.680	6.848	6.989	7.109	7.213
19	5.752	5.846	5.932	6.009	6.081	6.147	6.209	6.267	6.321	6.371	6.419	6.626	6.792	6.930	7.048	7.152
20	5.714	5.807	5.891	5.968	6.039	6.104	6.165	6.222	6.275	6.325	6.373	6.576	6.740	6.877	6.994	7.097
24	5.594	5.683	5.764	5.838	5.906	5.968	6.027	6.081	6.132	6.181	6.226	6.421	6.579	6.710	6.822	6.920
30	5.475	5.561	5.638	5.709	5.774	5.833	5.889	5.941	5.990	6.037	6.080	6.267	6.417	6.543	6.650	6.744
40	5.358	5.439	5.513	5.581	5.642	5.700	5.753	5.803	5.849	5.893	5.934	6.112	6.255	6.375	6.477	6.566
60	5.241	5.319	5.389	5.453	5.512	5.566	5.617	5.664	5.708	5.750	5.789	5.958	6.093	6.206	6.303	6.387
120	5.126	5.200	5.266	5.327	5.382	5.434	5.481	5.526	5.568	5.607	5.644	5.802	5.929	6.035	6.126	6.205
∞	5.012	5.081	5.144	5.201	5.253	5.301	5.346	5.388	5.427	5.463	5.498	5.646	5.764	5.863	5.947	6.020

$\alpha = 0.01$



k	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	90.03	135.0	164.3	185.6	202.2	215.8	227.2	237.0	245.6	253.2	260.0	266.2	271.8	277.0	281.8	286.3	290.4
2	14.04	19.02	22.29	24.72	26.63	28.20	29.53	30.68	31.69	32.59	33.40	34.13	34.81	35.43	36.00	36.53	37.03
3	8.261	10.62	12.17	13.33	14.24	15.00	15.64	16.20	16.69	17.13	17.53	17.89	18.22	18.52	18.81	19.07	19.32
4	6.512	8.120	9.173	9.958	10.58	11.10	11.55	11.93	12.27	12.57	12.84	13.09	13.32	13.53	13.73	13.91	14.08
5	5.702	6.976	7.804	8.421	8.913	9.321	9.669	9.972	10.24	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68
6	5.243	6.331	7.033	7.556	7.973	8.318	8.613	8.869	9.097	9.301	9.485	9.653	9.808	9.951	10.08	10.21	10.32
7	4.949	5.919	6.543	7.005	7.373	7.679	7.939	8.166	8.368	8.548	8.711	8.860	8.997	9.124	9.242	9.353	9.456
8	4.746	5.635	6.204	6.625	6.960	7.237	7.474	7.681	7.863	8.027	8.176	8 312	8.436	8.552	8.659	8.760	8.854
9	4.596	5.428	5.957	6.348	6.658	6.915	7.134	7.325	7.495	7.647	7.784	7.910	8.025	8.132	8.232	8.325	8.412
10	4.482	5.270	5.769	6.136	6.428	6.669	6.875	7.055	7.213	7.356	7.485	7.603	7.712	7.812	7.906	7.993	8.076
11	4.392	5.146	5.621	5.970	6.247	6.476	6.672	6.842	6.992	7.128	7 250	7.362	7.465	7.560	7.649	7.732	7.809
12	4.320	5.046	5.502	5.836	6.101	6.321	6.507	6.670	6.814	6.943	7.060	7.167	7.265	7.356	7.441	7.520	7.594
13	4.260	4.964	5.404	5.727	5.981	6.192	6.372	6.528	6.667	6.791	6.903	7.006	7.101	7.188	7.269	7 345	7.417
14	4.210	4.895	5.322	5.634	5.881	6.085	6.258	6.409	6.543	6.664	6.772	6.871	6.962	7.047	7.126	7.199	7.268
15	4.168	4.836	5.252	5.556	5.796	5.994	6.162	6.309	6.439	6.555	6 660	6.757	6.845	6.927	7.003	7.074	7.142
16	4.131	4.786	5.192	5.489	5.722	5.915	6.079	6.222	6.349	6.462	6.564	6.658	6.744	6.823	6.898	6.967	7.032
17	4.099	4.742	5.140	5.430	5.659	5.847	6.007	6.147	6.270	6.381	6.480	6.572	6.656	6.734	6.806	6.873	6.937
18	4.071	4.703	5.094	5.379	5.603	5.788	5.944	6.081	6.201	6.310	6.407	6.497	6.579	6.655	6.725	6.792	6.854
19	4.046	4.670	5.054	5.334	5.554	5.735	5.889	6.022	6.141	6.247	6.342	6.430	6.510	6.586	6.654	6.719	6.780
20	4.024	4.639	5.018	5.294	5.510	5.688	5.839	5.970	6.087	6.191	6.285	6.371	6.450	6.523	6.591	6.654	6.714
24	3.956	4.546	4.907	5.168	5.374	5.542	5.685	5.809	5.919	6.017	6.106	6.186	6.261	6.330	6.394	6.453	6.510
30	3.889	4.455	4.799	5.048	5.242	5.401	5.536	5.653	5.756	5.849	5.932	6.008	6.078	6.143	6.203	6.259	6.311
40	3.825	4.367	4.696	4.931	5.114	5.265	5.392	5.502	5.599	5.686	5.764	5.835	5.900	5.961	6.017	6.069	6.119
60	3.762	4.282	4.595	4.818	4.991	5.133	5.253	5.356	5.447	5.528	5.601	5.667	5.728	5.785	5.837	5.886	5.931
120	3.702	4.200	4.497	4.709	4.872	5.005	5.118	5.214	5.299	5.375	5.443	5.505	5.562	5 614	5.662	5.708	5.750
∞	3.643	4.120	4.403	4.603	4.757	4.882	4.987	5.078	5.157	5.227	5.290	5.348	5.400	5.448	5.493	5.535	5.574

k	20	22	24	26	28	30	32	34	36	38	40	50	60	70	80	90
1	298.0	304.7	310.8	316.3	321.3	326.0	330.3	334.3	338.0	341.5	344 8	358.9	370.1	379 4	387.3	394.1
2	37.95	38.76	39.49	40.15	40.76	41.32	41.84	42.33	42.78	43.21	43.61	45.33	46.70	47.83	48.80	49.64
3	19.77	20.17	20.53	20.86	21.16	21.44	21.70	21.95	22.17	22.39	22.59	23.45	24 13	24.71	25.19	25.62
4	14.40	14.68	14.93	15.16	15.37	15.57	15.75	15.92	16.08	16.23	16.37	16.98	17.46	17.86	18.20	18.50
5	11.93	12.16	12.36	12.54	12.71	12.87	13.02	13.15	13.28	13.40	13.52	14.00	14.39	14.72	14.99	15.23
6	10.54	10.73	10.91	11.06	11.21	11.34	11.47	11.58	11.69	11.80	11.90	12.31	12.65	12.92	13.16	13.37
7	9.646	9.815	9.970	10.11	10.24	10.36	10.47	10.58	10.67	10.77	10.85	11.23	11.52	11.77	11.99	12.17
8	9.027	9.182	9.322	9.450	9.569	9.678	9.779	9.874	9.964	10.05	10.13	10.47	10.75	10.97	11.17	11.34
9	8.573	8.717	8.847	8.966	9.075	9.177	9.271	9.360	9.443	9.521	9.594	9.912	10.17	10.38	10.57	10.73
10	8.226	8.361	8.483	8.595	8.698	8.794	8.883	8.966	9.044	9.117	9.187	9.486	9.726	9.927	10.10	10.25
11	7.952	8.080	8.196	8.303	8.400	8.491	8.575	8.654	8.728	8.798	8.864	9.148	9.377	9.568	9.732	9.875
12	7.731	7.853	7.964	8.066	8.159	8.246	8.327	8.402	8.473	8.539	8.603	8.875	9.094	9.277	9.434	9.571
13	7.548	7.665	7.772	7.870	7.960	8.043	8.121	8.193	8.262	8.326	8.387	8.648	8.859	9.035	9.187	9.318
14	7.395	7.508	7.611	7.705	7.792	7.873	7.948	8.018	8.084	8.146	8.204	8.457	8.661	8.832	8.978	9.106
15	7.264	7.374	7.474	7.566	7.650	7.728	7.800	7.869	7.932	7.992	8.049	8.295	8.492	8.658	8.800	8.924
16	7.152	7.258	7.356	7.445	7.527	7.602	7.673	7.739	7.802	7.860	7 916	8.154	8.347	8.507	8.646	8.767
17	7.053	7.158	7.253	7.340	7.420	7.493	7.563	7.627	7.687	7 745	7.799	8.031	8.219	8.377	8.511	8.630
18	6.968	7.070	7.163	7.247	7.325	7.398	7.465	7.528	7.587	7.643	7.696	7.924	8 107	8.261	8.393	8.508
19	6.891	6.992	7.082	7.166	7.242	7.313	7.379	7.440	7.498	7.553	7.605	7.828	8.008	8.159	8.288	8.401
20	6.823	6.922	7.011	7.092	7.168	7.237	7.302	7.362	7.419	7.473	7.523	7.742	7.919	8.067	8.194	8.305
24	6.612	6.705	6.789	6.865	6.936	7.001	7.062	7.119	7.173	7.223	7.270	7.476	7.642	7.780	7.900	8.004
30	6.407	6.494	6.572	6.644	6.710	6.772	6.828	6.881	6.932	6.978	7 023	7.215	7.370	7.500	7.611	7.709
40	6.209	6.289	6.362	6.429	6.490	6.547	6.600	6.650	6.697	6.740	6.782	6.960	7.104	7.225	7.323	7.419
60	6.015	6.090	6.158	6.220	6.277	6.330	6.378	6.424	6.467	6.507	6.546	6.710	6.843	6.954	7.050	7.133
120	5.827	5.897	5.959	6.016	6.069	6.117	6.162	6.204	6.244	6.281	6.316	6.467	6.588	6.689	6.776	6.862
∞	5.645	5.709	5.766	5.818	5.866	5.911	5.952	5.990	6.026	6.060	6.092	6.228	6.338	6.429	6.507	6.575

ource: Adapted from Harter, H. L. (1960). "Tables of Range and Studentized Range," *Annals of Mathematical Statistics*, **31**, 1122–1147. Used by permission of th f Mathematical Statistics.

Source: Mason et al. (1989).

Table B-23.
Critical Values of Student's t-Distribution

p								
df	**0.8**	**0.9**	**0.95**	**0.975**	**0.99**	**0.995**	**0.999**	**0.9995**
1	1.376	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	1.061	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	0.9785	1.638	2.353	3.182	4.541	5.841	10.21	12.92
4	0.9410	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.9195	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.9057	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.8960	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.8889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.8834	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.8791	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.8755	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.8726	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.8702	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.8681	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.8662	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.8647	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.8633	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.8620	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.8610	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.8600	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.8591	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.8583	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.8575	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.8569	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.8562	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.8557	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.8551	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.8546	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.8542	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.8538	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.8507	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.8477	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.8446	1.289	1.658	1.980	2.358	2.617	3.160	3.373
	0.8417	1.282	1.645	1.960	2.327	2.576	3.091	3.291

NOTE: Table generated using SAS, a statistical software package. The percentiles $t_{p,\nu}$ are listed for various values of degrees of freedom (df), ν: $p = P(t_\nu \leq t_{p,\nu})$.

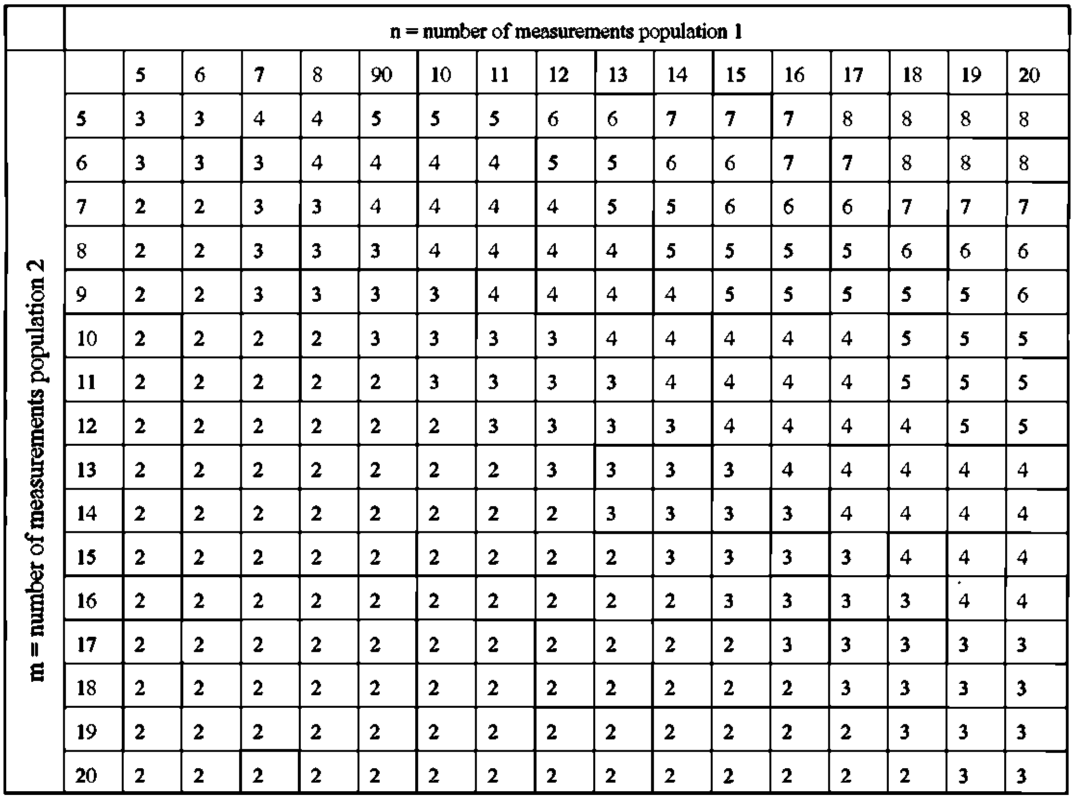
Table B-24.
Quantiles of the Wilcoxon Signed Rank Test

n	$w_{0.01}$	$w_{0.05}$	$w_{0.10}$	$w_{0.20}$
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

Source: EPA/600/R-96/084.

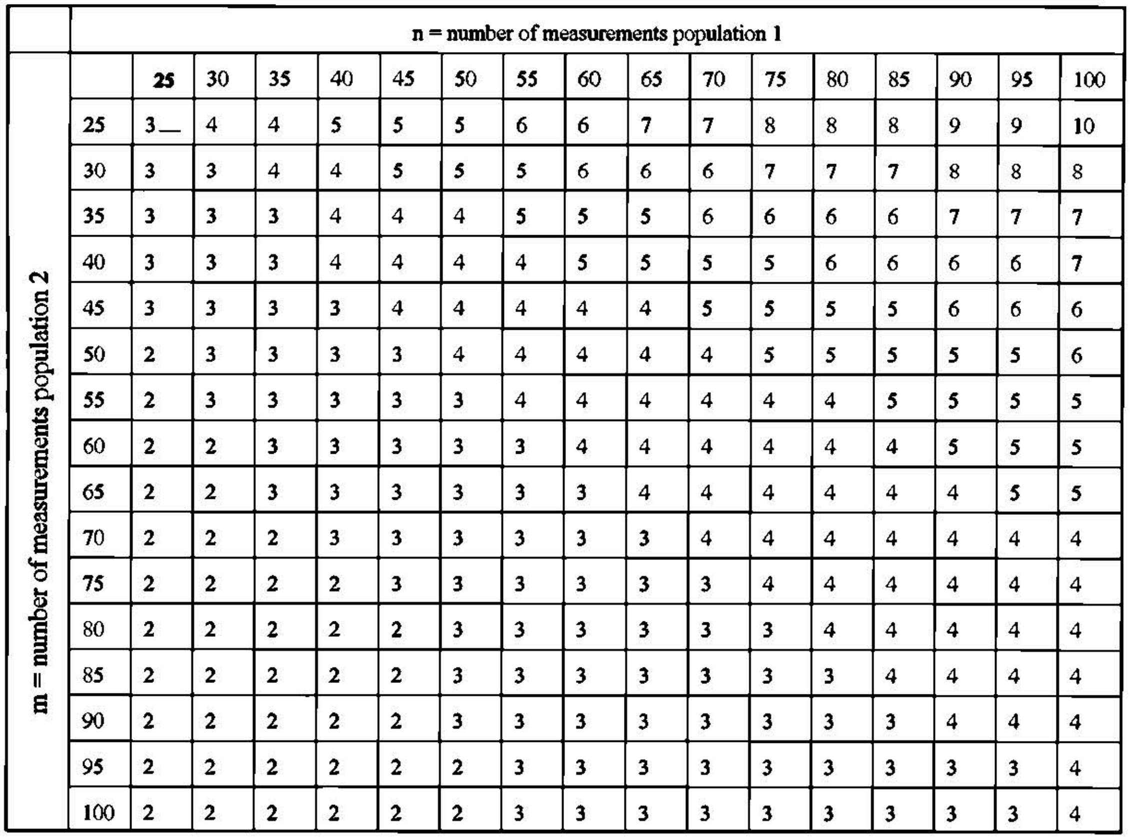
Table B-25.
Modified Quantile Test Critical Numbers Level of Significance (α)

For Approximately $\alpha = 0.10$



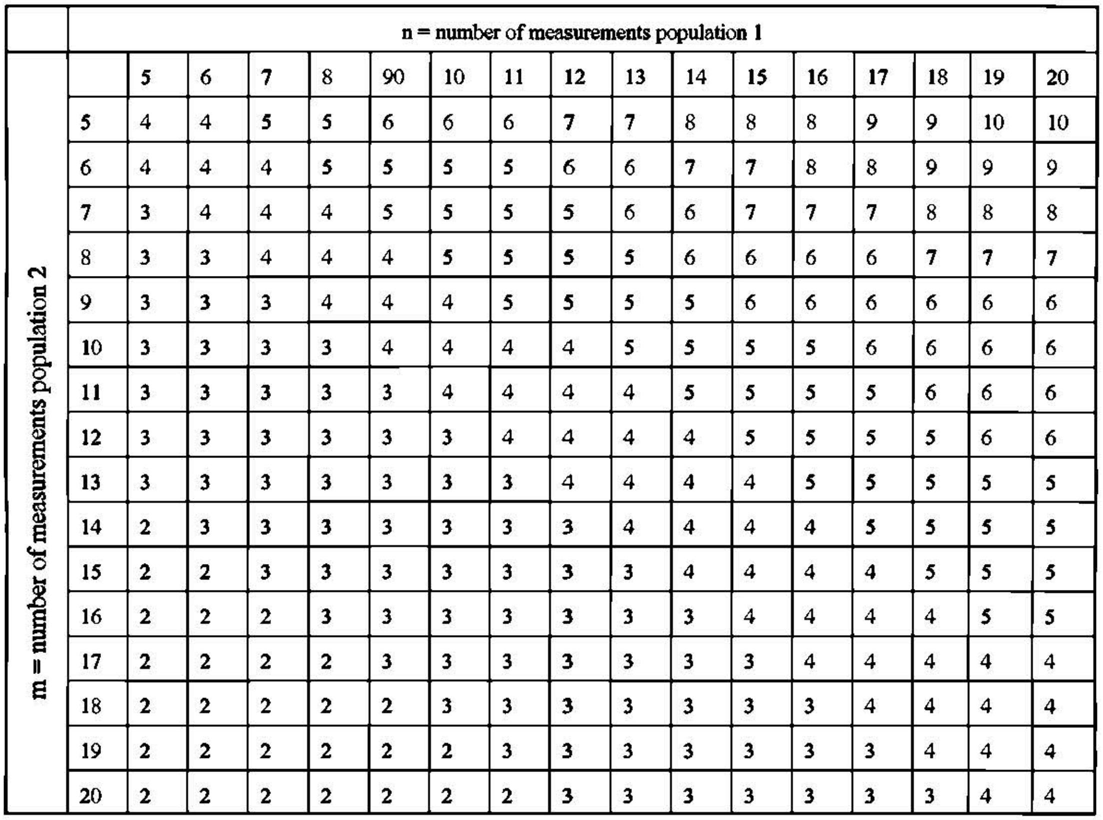
		n = number of measurements population 1															
m = number of measurements population 2		5	6	7	8	90	10	11	12	13	14	15	16	17	18	19	20
	5	3	3	4	4	5	5	5	6	6	7	7	7	8	8	8	8
	6	3	3	3	4	4	4	4	5	5	6	6	7	7	8	8	8
	7	2	2	3	3	4	4	4	4	5	5	6	6	6	7	7	7
	8	2	2	3	3	3	4	4	4	4	5	5	5	5	6	6	6
	9	2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	6
	10	2	2	2	2	3	3	3	3	4	4	4	4	4	5	5	5
	11	2	2	2	2	2	3	3	3	3	4	4	4	4	5	5	5
	12	2	2	2	2	2	2	3	3	3	3	4	4	4	4	5	5
	13	2	2	2	2	2	2	2	3	3	3	3	4	4	4	4	4
	14	2	2	2	2	2	2	2	2	3	3	3	3	4	4	4	4
	15	2	2	2	2	2	2	2	2	2	3	3	3	3	4	4	4
	16	2	2	2	2	2	2	2	2	2	2	3	3	3	3	4	4
	17	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3
	18	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3
	19	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3
	20	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3

For Approximately $\alpha = 0.10$



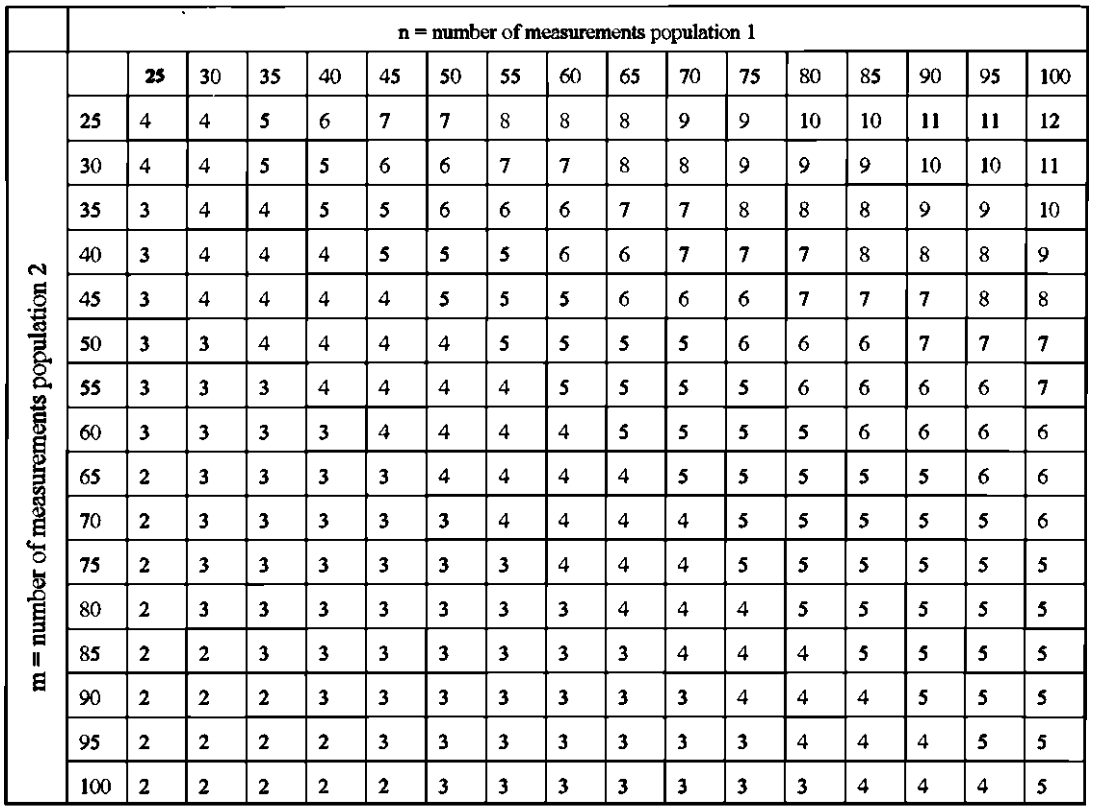
	n = number of measurements population 1																
m = number of measurements population 2		25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
	25	3__	4	4	5	5	5	6	6	7	7	8	8	8	9	9	10
	30	3	3	4	4	5	5	5	6	6	6	7	7	7	8	8	8
	35	3	3	3	4	4	4	5	5	5	6	6	6	6	7	7	7
	40	3	3	3	4	4	4	4	5	5	5	5	6	6	6	6	7
	45	3	3	3	3	4	4	4	4	4	5	5	5	5	6	6	6
	50	2	3	3	3	3	4	4	4	4	4	5	5	5	5	5	6
	55	2	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5
	60	2	2	3	3	3	3	3	4	4	4	4	4	4	5	5	5
	65	2	2	3	3	3	3	3	3	4	4	4	4	4	4	5	5
	70	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	4
	75	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
	80	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4
	85	2	2	2	2	2	3	3	3	3	3	3	3	4	4	4	4
	90	2	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4
	95	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4
	100	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	4

For Approximately $\alpha = 0.05$



	n = number of measurements population 1																
m = number of measurements population 2		5	6	7	8	90	10	11	12	13	14	15	16	17	18	19	20
	5	4	4	5	5	6	6	6	7	7	8	8	8	9	9	10	10
	6	4	4	4	5	5	5	5	6	6	7	7	8	8	9	9	9
	7	3	4	4	4	5	5	5	5	6	6	7	7	7	8	8	8
	8	3	3	4	4	4	5	5	5	5	6	6	6	6	7	7	7
	9	3	3	3	4	4	4	5	5	5	5	6	6	6	6	6	6
	10	3	3	3	3	4	4	4	4	5	5	5	5	6	6	6	6
	11	3	3	3	3	3	4	4	4	4	5	5	5	5	6	6	6
	12	3	3	3	3	3	3	4	4	4	4	5	5	5	5	6	6
	13	3	3	3	3	3	3	3	4	4	4	4	5	5	5	5	5
	14	2	3	3	3	3	3	3	3	4	4	4	4	5	5	5	5
	15	2	2	3	3	3	3	3	3	3	4	4	4	4	5	5	5
	16	2	2	2	3	3	3	3	3	3	3	4	4	4	4	5	5
	17	2	2	2	2	3	3	3	3	3	3	3	4	4	4	4	4
	18	2	2	2	2	2	3	3	3	3	3	3	3	4	4	4	4
	19	2	2	2	2	2	2	3	3	3	3	3	3	3	4	4	4
	20	2	2	2	2	2	2	2	3	3	3	3	3	3	3	4	4

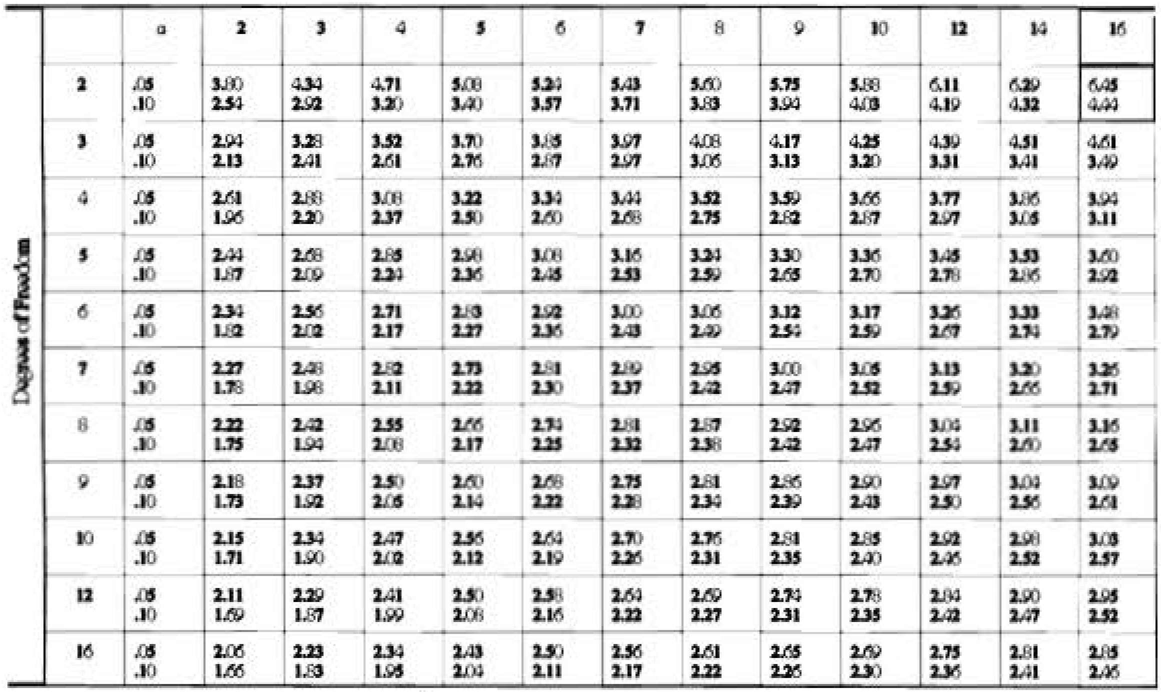
For Approximately $\alpha = 0.05$



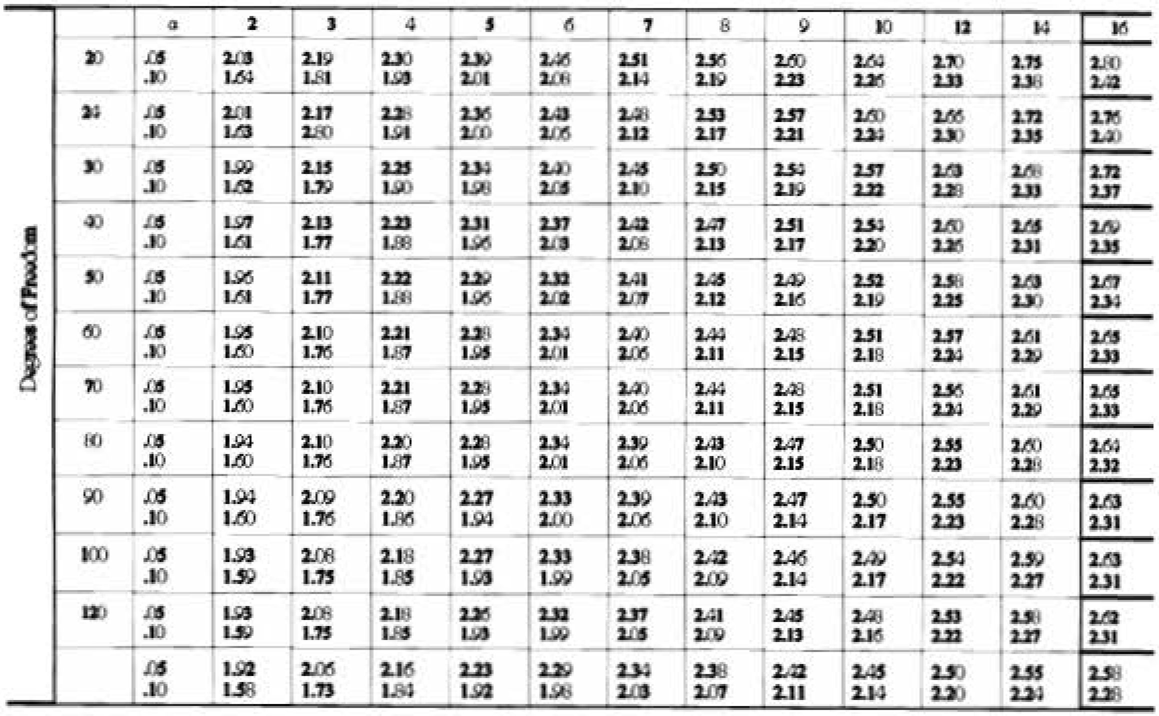
		n = number of measurements population 1															
m = number of measurements population 2		25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
	25	4	4	5	6	7	7	8	8	8	9	9	10	10	11	11	12
	30	4	4	5	5	6	6	7	7	8	8	9	9	9	10	10	11
	35	3	4	4	5	5	6	6	6	7	7	8	8	8	9	9	10
	40	3	4	4	4	5	5	5	6	6	7	7	7	8	8	8	9
	45	3	4	4	4	4	5	5	5	6	6	6	7	7	7	8	8
	50	3	3	4	4	4	4	5	5	5	5	6	6	6	7	7	7
	55	3	3	3	4	4	4	4	5	5	5	5	6	6	6	6	7
	60	3	3	3	3	4	4	4	4	5	5	5	5	6	6	6	6
	65	2	3	3	3	3	4	4	4	4	5	5	5	5	5	6	6
	70	2	3	3	3	3	3	4	4	4	4	5	5	5	5	5	6
	75	2	3	3	3	3	3	3	4	4	4	5	5	5	5	5	5
	80	2	3	3	3	3	3	3	3	4	4	4	5	5	5	5	5
	85	2	2	3	3	3	3	3	3	3	4	4	4	5	5	5	5
	90	2	2	2	3	3	3	3	3	3	3	4	4	4	5	5	5
	95	2	2	2	2	3	3	3	3	3	3	3	4	4	4	5	5
	100	2	2	2	2	2	3	3	3	3	3	3	3	4	4	4	5

Source: EPA/600/R-96/084.

Table B-26.
Dunnett's Test (One-Tailed) Total Number of Investigate Groups (*K* - 1)



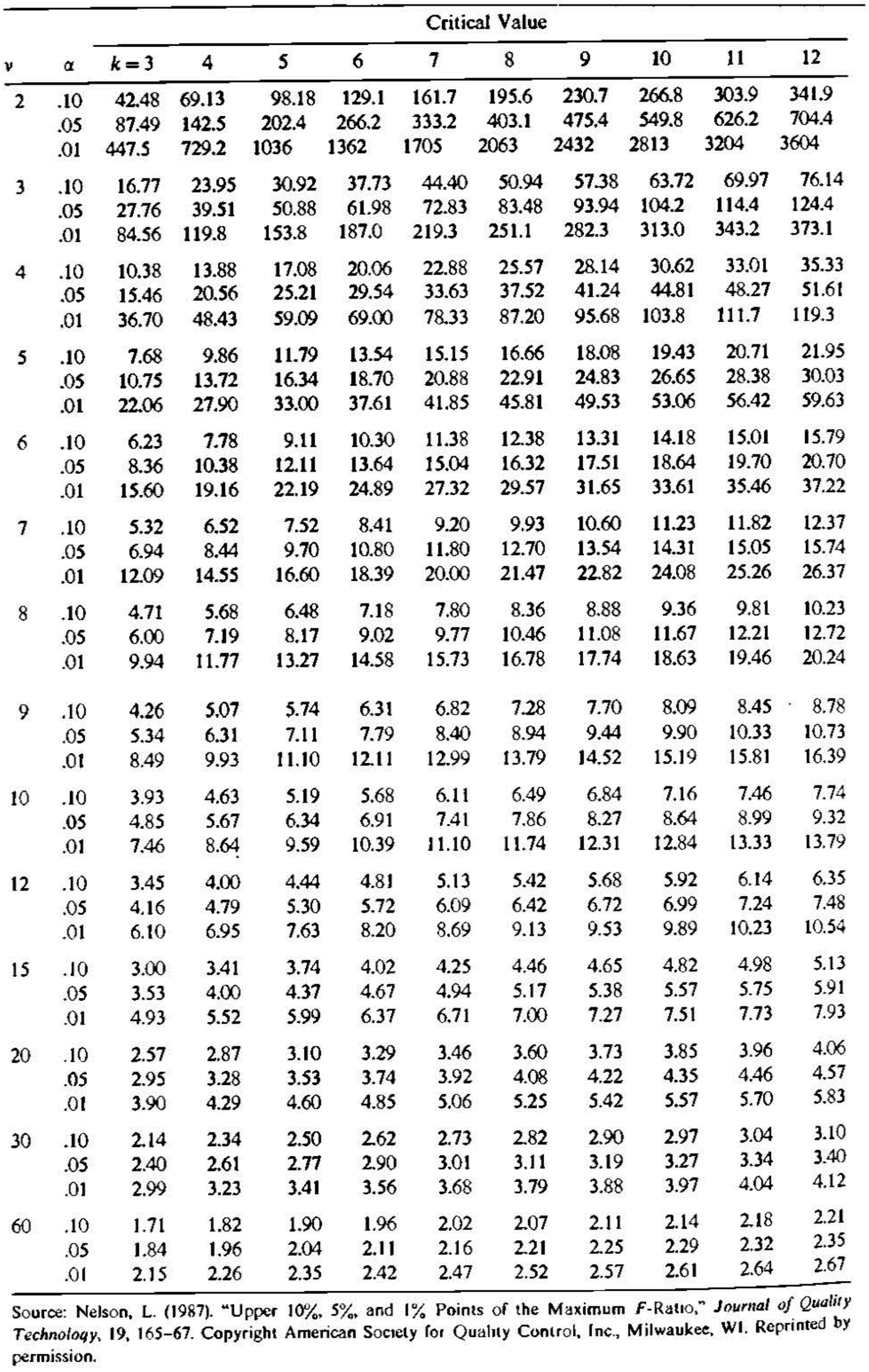
Degrees of Freedom	α	2	3	4	5	6	7	8	9	10	12	14	16
2	.05 .10	3.80 2.54	4.34 2.92	4.71 3.20	5.08 3.40	5.24 3.57	5.43 3.71	5.60 3.83	5.75 3.94	5.88 4.03	6.11 4.19	6.29 4.32	6.45 4.44
3	.05 .10	2.94 2.13	3.28 2.41	3.52 2.61	3.70 2.76	3.85 2.87	3.97 2.97	4.08 3.06	4.17 3.13	4.25 3.20	4.39 3.31	4.51 3.41	4.61 3.49
4	.05 .10	2.61 1.96	2.88 2.20	3.08 2.37	3.22 2.50	3.34 2.60	3.44 2.68	3.52 2.75	3.59 2.82	3.66 2.87	3.77 2.97	3.86 3.05	3.94 3.11
5	.05 .10	2.44 1.87	2.68 2.09	2.85 2.24	2.98 2.36	3.08 2.45	3.16 2.53	3.24 2.59	3.30 2.65	3.36 2.70	3.45 2.78	3.53 2.86	3.60 2.92
6	.05 .10	2.34 1.82	2.56 2.02	2.71 2.17	2.83 2.27	2.92 2.36	3.00 2.43	3.06 2.49	3.12 2.54	3.17 2.59	3.26 2.67	3.33 2.74	3.39 2.79
7	.05 .10	2.27 1.78	2.48 1.98	2.62 2.11	2.73 2.22	2.81 2.30	2.89 2.37	2.95 2.42	3.00 2.47	3.05 2.52	3.13 2.59	3.20 2.66	3.26 2.71
8	.05 .10	2.22 1.75	2.42 1.94	2.55 2.08	2.66 2.17	2.74 2.25	2.81 2.32	2.87 2.38	2.92 2.42	2.96 2.47	3.04 2.54	3.11 2.60	3.16 2.65
9	.05 .10	2.18 1.73	2.37 1.92	2.50 2.05	2.60 2.14	2.68 2.22	2.75 2.28	2.81 2.34	2.86 2.39	2.90 2.43	2.97 2.50	3.04 2.56	3.09 2.61
10	.05 .10	2.15 1.71	2.34 1.90	2.47 2.02	2.56 2.12	2.64 2.19	2.70 2.26	2.76 2.31	2.81 2.35	2.85 2.40	2.92 2.46	2.98 2.52	3.03 2.57
12	.05 .10	2.11 1.69	2.29 1.87	2.41 1.99	2.50 2.08	2.58 2.16	2.64 2.22	2.69 2.27	2.74 2.31	2.78 2.35	2.84 2.42	2.90 2.47	2.95 2.52
16	.05 .10	2.06 1.66	2.23 1.83	2.34 1.95	2.43 2.04	2.50 2.11	2.56 2.17	2.61 2.22	2.65 2.26	2.69 2.30	2.75 2.36	2.81 2.41	2.85 2.46



Degrees of Freedom	α	2	3	4	5	6	7	8	9	10	12	14	16
20	.05 .10	2.03 1.64	2.19 1.81	2.30 1.93	2.39 2.01	2.46 2.08	2.51 2.14	2.56 2.19	2.60 2.23	2.64 2.26	2.70 2.33	2.75 2.38	2.80 2.42
24	.05 .10	2.01 1.63	2.17 2.80	2.28 1.91	2.36 2.00	2.43 2.05	2.48 2.12	2.53 2.17	2.57 2.21	2.60 2.24	2.66 2.30	2.72 2.35	2.76 2.40
30	.05 .10	1.99 1.62	2.15 1.79	2.25 1.90	2.34 1.98	2.40 2.05	2.45 2.10	2.50 2.15	2.54 2.19	2.57 2.22	2.63 2.28	2.68 2.33	2.72 2.37
40	.05 .10	1.97 1.61	2.13 1.77	2.23 1.88	2.31 1.96	2.37 2.03	2.42 2.08	2.47 2.13	2.51 2.17	2.54 2.20	2.60 2.26	2.65 2.31	2.69 2.35
50	.05 .10	1.96 1.61	2.11 1.77	2.22 1.88	2.29 1.96	2.32 2.02	2.41 2.07	2.45 2.12	2.49 2.16	2.52 2.19	2.58 2.25	2.63 2.30	2.67 2.34
60	.05 .10	1.95 1.60	2.10 1.76	2.21 1.87	2.28 1.95	2.34 2.01	2.40 2.06	2.44 2.11	2.48 2.15	2.51 2.18	2.57 2.24	2.61 2.29	2.65 2.33
70	.05 .10	1.95 1.60	2.10 1.76	2.21 1.87	2.28 1.95	2.34 2.01	2.40 2.06	2.44 2.11	2.48 2.15	2.51 2.18	2.56 2.24	2.61 2.29	2.65 2.33
80	.05 .10	1.94 1.60	2.10 1.76	2.20 1.87	2.28 1.95	2.34 2.01	2.39 2.06	2.43 2.10	2.47 2.15	2.50 2.18	2.55 2.23	2.60 2.28	2.64 2.32
90	.05 .10	1.94 1.60	2.09 1.76	2.20 1.86	2.27 1.94	2.33 2.00	2.39 2.06	2.43 2.10	2.47 2.14	2.50 2.17	2.55 2.23	2.60 2.28	2.63 2.31
100	.05 .10	1.93 1.59	2.08 1.75	2.18 1.85	2.27 1.93	2.33 1.99	2.38 2.05	2.42 2.09	2.46 2.14	2.49 2.17	2.54 2.22	2.59 2.27	2.63 2.31
120	.05 .10	1.93 1.59	2.08 1.75	2.18 1.85	2.26 1.93	2.32 1.99	2.37 2.05	2.41 2.09	2.45 2.13	2.48 2.16	2.53 2.22	2.58 2.27	2.62 2.31
	.05 .10	1.92 1.58	2.06 1.73	2.16 1.84	2.23 1.92	2.29 1.98	2.34 2.03	2.38 2.07	2.42 2.11	2.45 2.14	2.50 2.20	2.55 2.24	2.58 2.28

Source: EPA/600/R-96/084.

Table B-27.
Upper Tail Critical Values for the *F*-Max Test

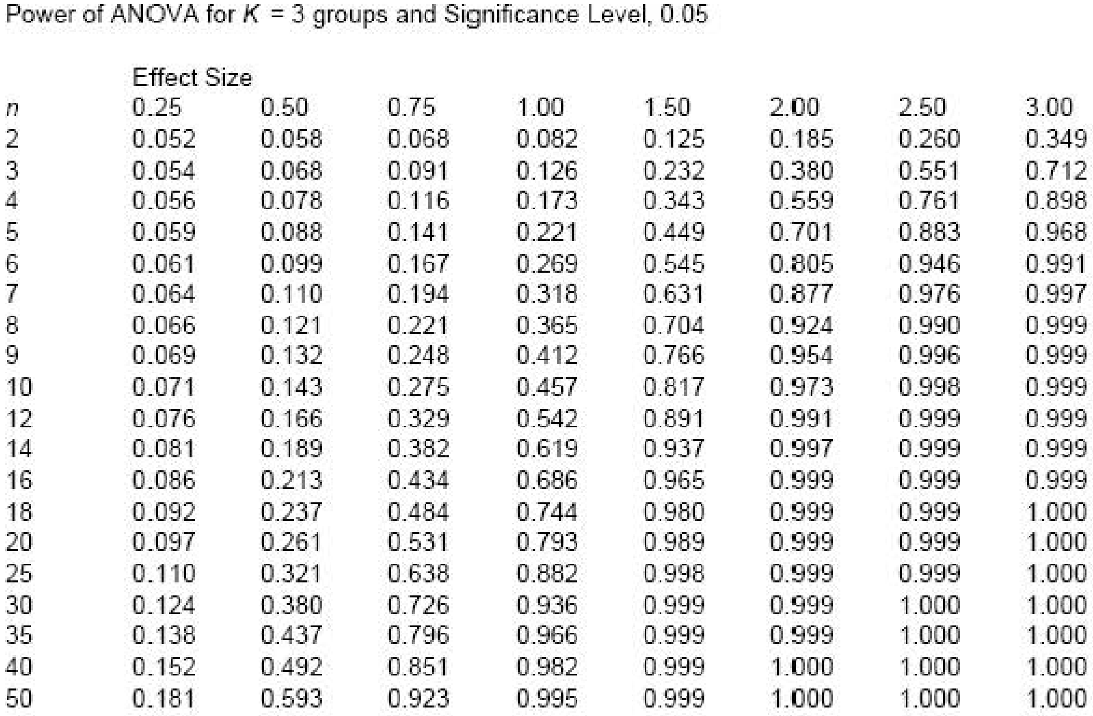


		Critical Value									
ν	α	k = 3	4	5	6	7	8	9	10	11	12
2	.10	42.48	69.13	98.18	129.1	161.7	195.6	230.7	266.8	303.9	341.9
	.05	87.49	142.5	202.4	266.2	333.2	403.1	475.4	549.8	626.2	704.4
	.01	447.5	729.2	1036	1362	1705	2063	2432	2813	3204	3604
3	.10	16.77	23.95	30.92	37.73	44.40	50.94	57.38	63.72	69.97	76.14
	.05	27.76	39.51	50.88	61.98	72.83	83.48	93.94	104.2	114.4	124.4
	.01	84.56	119.8	153.8	187.0	219.3	251.1	282.3	313.0	343.2	373.1
4	.10	10.38	13.88	17.08	20.06	22.88	25.57	28.14	30.62	33.01	35.33
	.05	15.46	20.56	25.21	29.54	33.63	37.52	41.24	44.81	48.27	51.61
	.01	36.70	48.43	59.09	69.00	78.33	87.20	95.68	103.8	111.7	119.3
5	.10	7.68	9.86	11.79	13.54	15.15	16.66	18.08	19.43	20.71	21.95
	.05	10.75	13.72	16.34	18.70	20.88	22.91	24.83	26.65	28.38	30.03
	.01	22.06	27.90	33.00	37.61	41.85	45.81	49.53	53.06	56.42	59.63
6	.10	6.23	7.78	9.11	10.30	11.38	12.38	13.31	14.18	15.01	15.79
	.05	8.36	10.38	12.11	13.64	15.04	16.32	17.51	18.64	19.70	20.70
	.01	15.60	19.16	22.19	24.89	27.32	29.57	31.65	33.61	35.46	37.22
7	.10	5.32	6.52	7.52	8.41	9.20	9.93	10.60	11.23	11.82	12.37
	.05	6.94	8.44	9.70	10.80	11.80	12.70	13.54	14.31	15.05	15.74
	.01	12.09	14.55	16.60	18.39	20.00	21.47	22.82	24.08	25.26	26.37
8	.10	4.71	5.68	6.48	7.18	7.80	8.36	8.88	9.36	9.81	10.23
	.05	6.00	7.19	8.17	9.02	9.77	10.46	11.08	11.67	12.21	12.72
	.01	9.94	11.77	13.27	14.58	15.73	16.78	17.74	18.63	19.46	20.24
9	.10	4.26	5.07	5.74	6.31	6.82	7.28	7.70	8.09	8.45	8.78
	.05	5.34	6.31	7.11	7.79	8.40	8.94	9.44	9.90	10.33	10.73
	.01	8.49	9.93	11.10	12.11	12.99	13.79	14.52	15.19	15.81	16.39
10	.10	3.93	4.63	5.19	5.68	6.11	6.49	6.84	7.16	7.46	7.74
	.05	4.85	5.67	6.34	6.91	7.41	7.86	8.27	8.64	8.99	9.32
	.01	7.46	8.64	9.59	10.39	11.10	11.74	12.31	12.84	13.33	13.79
12	.10	3.45	4.00	4.44	4.81	5.13	5.42	5.68	5.92	6.14	6.35
	.05	4.16	4.79	5.30	5.72	6.09	6.42	6.72	6.99	7.24	7.48
	.01	6.10	6.95	7.63	8.20	8.69	9.13	9.53	9.89	10.23	10.54
15	.10	3.00	3.41	3.74	4.02	4.25	4.46	4.65	4.82	4.98	5.13
	.05	3.53	4.00	4.37	4.67	4.94	5.17	5.38	5.57	5.75	5.91
	.01	4.93	5.52	5.99	6.37	6.71	7.00	7.27	7.51	7.73	7.93
20	.10	2.57	2.87	3.10	3.29	3.46	3.60	3.73	3.85	3.96	4.06
	.05	2.95	3.28	3.53	3.74	3.92	4.08	4.22	4.35	4.46	4.57
	.01	3.90	4.29	4.60	4.85	5.06	5.25	5.42	5.57	5.70	5.83
30	.10	2.14	2.34	2.50	2.62	2.73	2.82	2.90	2.97	3.04	3.10
	.05	2.40	2.61	2.77	2.90	3.01	3.11	3.19	3.27	3.34	3.40
	.01	2.99	3.23	3.41	3.56	3.68	3.79	3.88	3.97	4.04	4.12
60	.10	1.71	1.82	1.90	1.96	2.02	2.07	2.11	2.14	2.18	2.21
	.05	1.84	1.96	2.04	2.11	2.16	2.21	2.25	2.29	2.32	2.35
	.01	2.15	2.26	2.35	2.42	2.47	2.52	2.57	2.61	2.64	2.67

Source: Nelson, L. (1987). "Upper 10%, 5%, and 1% Points of the Maximum *F*-Ratio," *Journal of Quality Technology*, 19, 165–67. Copyright American Society for Quality Control, Inc., Milwaukee, WI. Reprinted by permission.

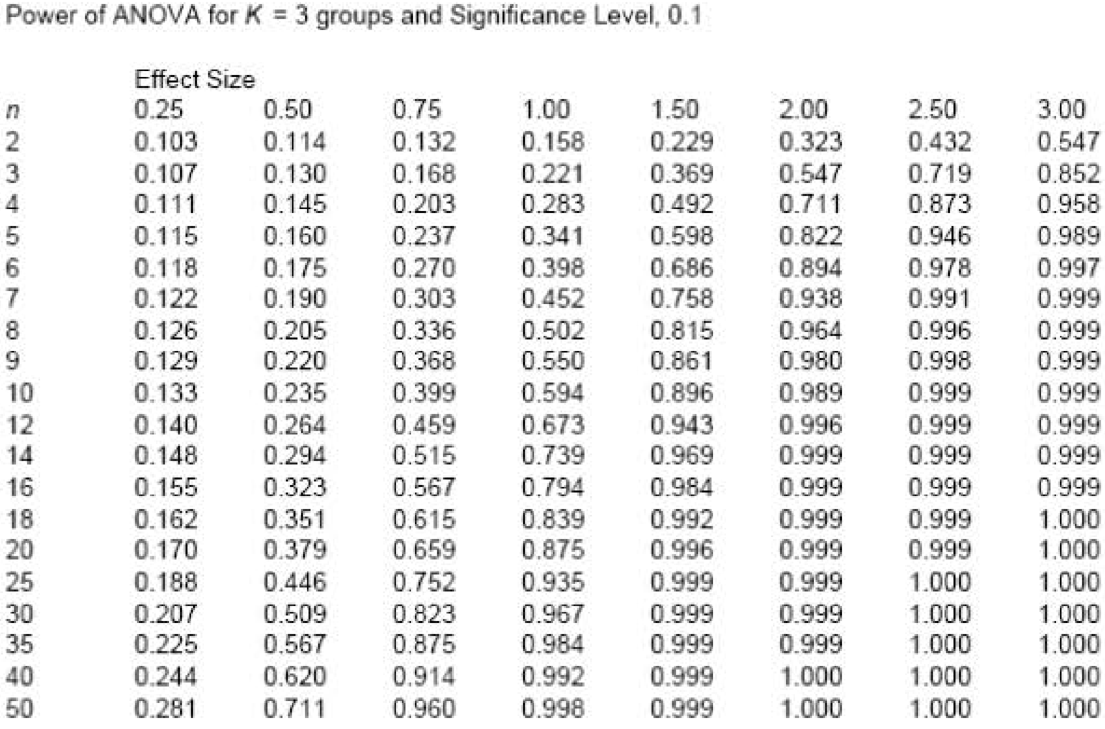
Source: Mason et al. (1989).

Table B-28.



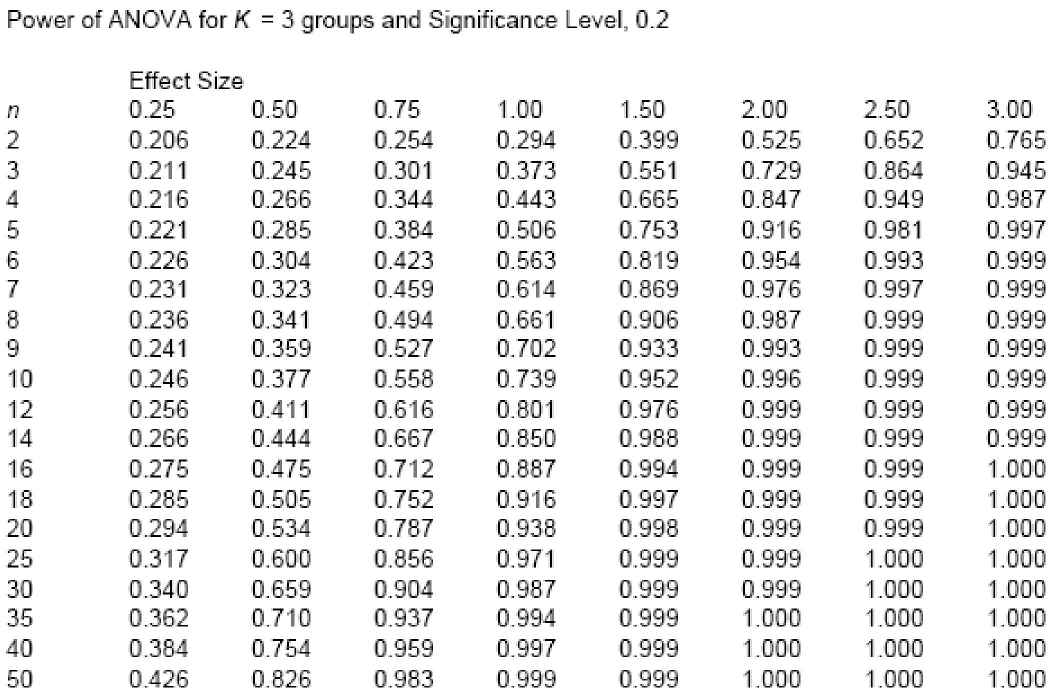
Power of ANOVA for *K* = 3 groups and Significance Level, 0.05

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.052	0.058	0.068	0.082	0.125	0.185	0.260	0.349
3	0.054	0.068	0.091	0.126	0.232	0.380	0.551	0.712
4	0.056	0.078	0.116	0.173	0.343	0.559	0.761	0.898
5	0.059	0.088	0.141	0.221	0.449	0.701	0.883	0.968
6	0.061	0.099	0.167	0.269	0.545	0.805	0.946	0.991
7	0.064	0.110	0.194	0.318	0.631	0.877	0.976	0.997
8	0.066	0.121	0.221	0.365	0.704	0.924	0.990	0.999
9	0.069	0.132	0.248	0.412	0.766	0.954	0.996	0.999
10	0.071	0.143	0.275	0.457	0.817	0.973	0.998	0.999
12	0.076	0.166	0.329	0.542	0.891	0.991	0.999	0.999
14	0.081	0.189	0.382	0.619	0.937	0.997	0.999	0.999
16	0.086	0.213	0.434	0.686	0.965	0.999	0.999	0.999
18	0.092	0.237	0.484	0.744	0.980	0.999	0.999	1.000
20	0.097	0.261	0.531	0.793	0.989	0.999	0.999	1.000
25	0.110	0.321	0.638	0.882	0.998	0.999	0.999	1.000
30	0.124	0.380	0.726	0.936	0.999	0.999	1.000	1.000
35	0.138	0.437	0.796	0.966	0.999	0.999	1.000	1.000
40	0.152	0.492	0.851	0.982	0.999	1.000	1.000	1.000
50	0.181	0.593	0.923	0.995	0.999	1.000	1.000	1.000



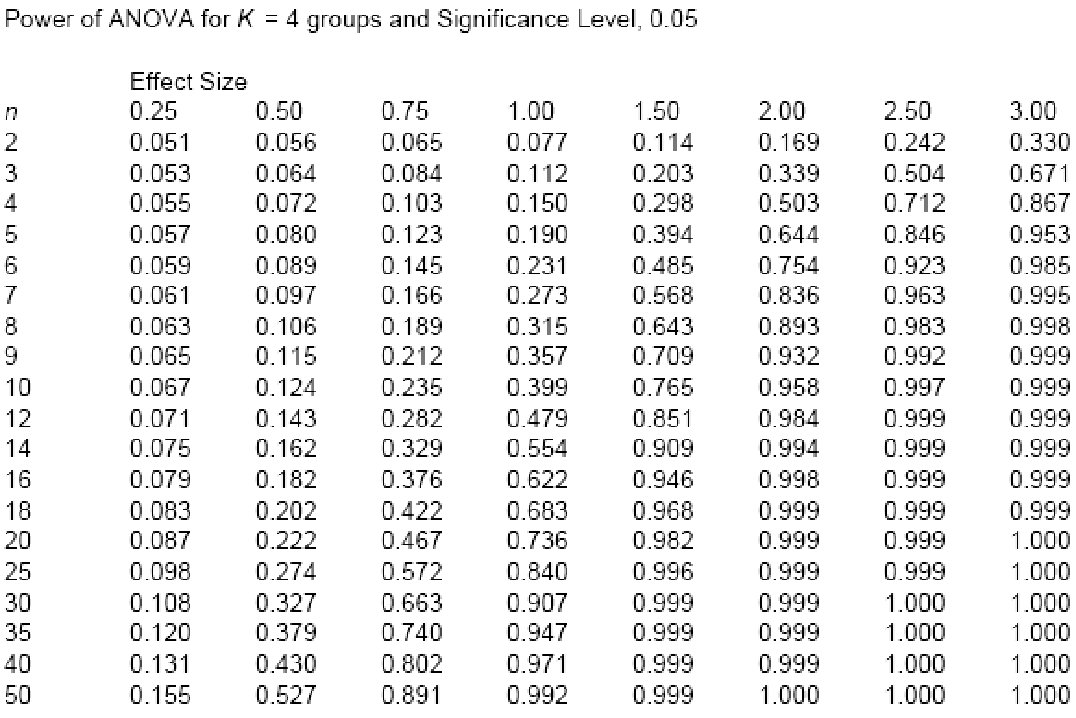
Power of ANOVA for *K* = 3 groups and Significance Level, 0.1

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.103	0.114	0.132	0.158	0.229	0.323	0.432	0.547
3	0.107	0.130	0.168	0.221	0.369	0.547	0.719	0.852
4	0.111	0.145	0.203	0.283	0.492	0.711	0.873	0.958
5	0.115	0.160	0.237	0.341	0.598	0.822	0.946	0.989
6	0.118	0.175	0.270	0.398	0.686	0.894	0.978	0.997
7	0.122	0.190	0.303	0.452	0.758	0.938	0.991	0.999
8	0.126	0.205	0.336	0.502	0.815	0.964	0.996	0.999
9	0.129	0.220	0.368	0.550	0.861	0.980	0.998	0.999
10	0.133	0.235	0.399	0.594	0.896	0.989	0.999	0.999
12	0.140	0.264	0.459	0.673	0.943	0.996	0.999	0.999
14	0.148	0.294	0.515	0.739	0.969	0.999	0.999	0.999
16	0.155	0.323	0.567	0.794	0.984	0.999	0.999	0.999
18	0.162	0.351	0.615	0.839	0.992	0.999	0.999	1.000
20	0.170	0.379	0.659	0.875	0.996	0.999	0.999	1.000
25	0.188	0.446	0.752	0.935	0.999	0.999	1.000	1.000
30	0.207	0.509	0.823	0.967	0.999	0.999	1.000	1.000
35	0.225	0.567	0.875	0.984	0.999	0.999	1.000	1.000
40	0.244	0.620	0.914	0.992	0.999	1.000	1.000	1.000
50	0.281	0.711	0.960	0.998	0.999	1.000	1.000	1.000



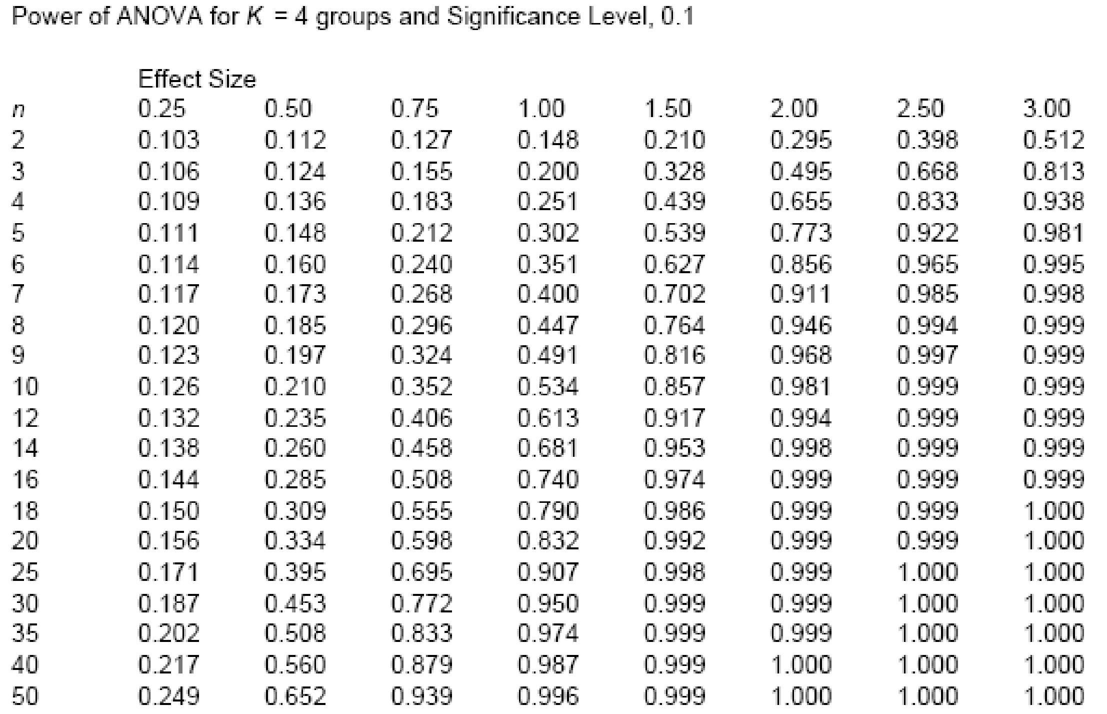
Power of ANOVA for *K* = 3 groups and Significance Level, 0.2

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.206	0.224	0.254	0.294	0.399	0.525	0.652	0.765
3	0.211	0.245	0.301	0.373	0.551	0.729	0.864	0.945
4	0.216	0.266	0.344	0.443	0.665	0.847	0.949	0.987
5	0.221	0.285	0.384	0.506	0.753	0.916	0.981	0.997
6	0.226	0.304	0.423	0.563	0.819	0.954	0.993	0.999
7	0.231	0.323	0.459	0.614	0.869	0.976	0.997	0.999
8	0.236	0.341	0.494	0.661	0.906	0.987	0.999	0.999
9	0.241	0.359	0.527	0.702	0.933	0.993	0.999	0.999
10	0.246	0.377	0.558	0.739	0.952	0.996	0.999	0.999
12	0.256	0.411	0.616	0.801	0.976	0.999	0.999	0.999
14	0.266	0.444	0.667	0.850	0.988	0.999	0.999	0.999
16	0.275	0.475	0.712	0.887	0.994	0.999	0.999	1.000
18	0.285	0.505	0.752	0.916	0.997	0.999	0.999	1.000
20	0.294	0.534	0.787	0.938	0.998	0.999	0.999	1.000
25	0.317	0.600	0.856	0.971	0.999	0.999	1.000	1.000
30	0.340	0.659	0.904	0.987	0.999	0.999	1.000	1.000
35	0.362	0.710	0.937	0.994	0.999	1.000	1.000	1.000
40	0.384	0.754	0.959	0.997	0.999	1.000	1.000	1.000
50	0.426	0.826	0.983	0.999	0.999	1.000	1.000	1.000



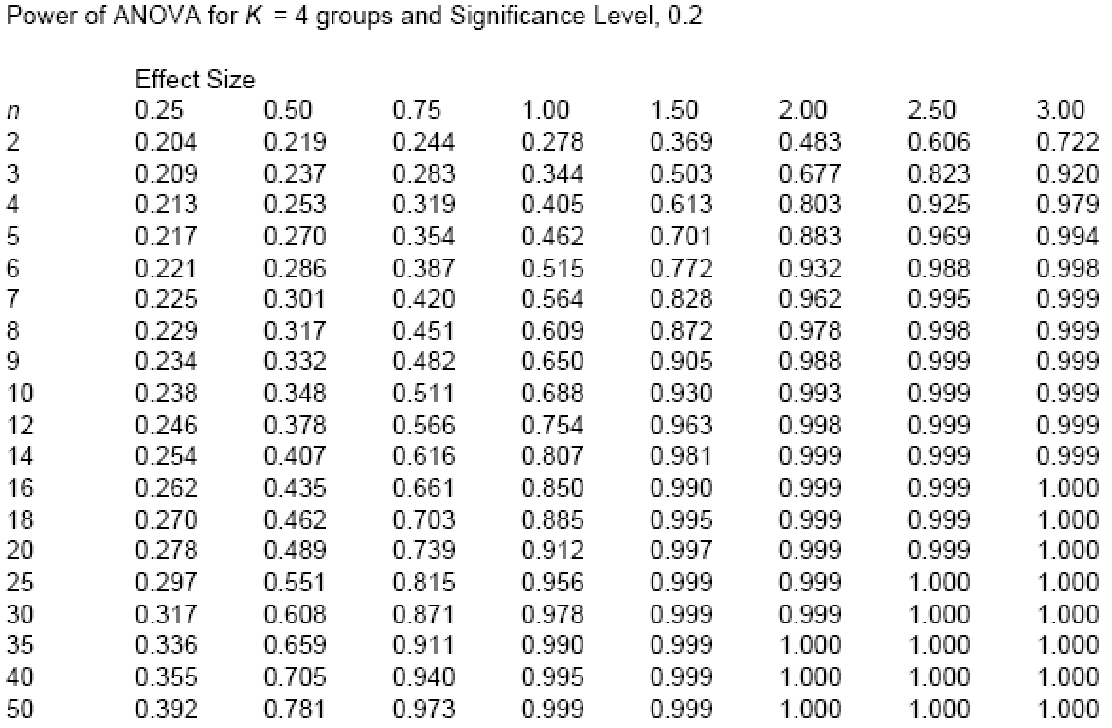
Power of ANOVA for *K* = 4 groups and Significance Level, 0.05

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.051	0.056	0.065	0.077	0.114	0.169	0.242	0.330
3	0.053	0.064	0.084	0.112	0.203	0.339	0.504	0.671
4	0.055	0.072	0.103	0.150	0.298	0.503	0.712	0.867
5	0.057	0.080	0.123	0.190	0.394	0.644	0.846	0.953
6	0.059	0.089	0.145	0.231	0.485	0.754	0.923	0.985
7	0.061	0.097	0.166	0.273	0.568	0.836	0.963	0.995
8	0.063	0.106	0.189	0.315	0.643	0.893	0.983	0.998
9	0.065	0.115	0.212	0.357	0.709	0.932	0.992	0.999
10	0.067	0.124	0.235	0.399	0.765	0.958	0.997	0.999
12	0.071	0.143	0.282	0.479	0.851	0.984	0.999	0.999
14	0.075	0.162	0.329	0.554	0.909	0.994	0.999	0.999
16	0.079	0.182	0.376	0.622	0.946	0.998	0.999	0.999
18	0.083	0.202	0.422	0.683	0.968	0.999	0.999	0.999
20	0.087	0.222	0.467	0.736	0.982	0.999	0.999	1.000
25	0.098	0.274	0.572	0.840	0.996	0.999	0.999	1.000
30	0.108	0.327	0.663	0.907	0.999	0.999	1.000	1.000
35	0.120	0.379	0.740	0.947	0.999	0.999	1.000	1.000
40	0.131	0.430	0.802	0.971	0.999	0.999	1.000	1.000
50	0.155	0.527	0.891	0.992	0.999	1.000	1.000	1.000



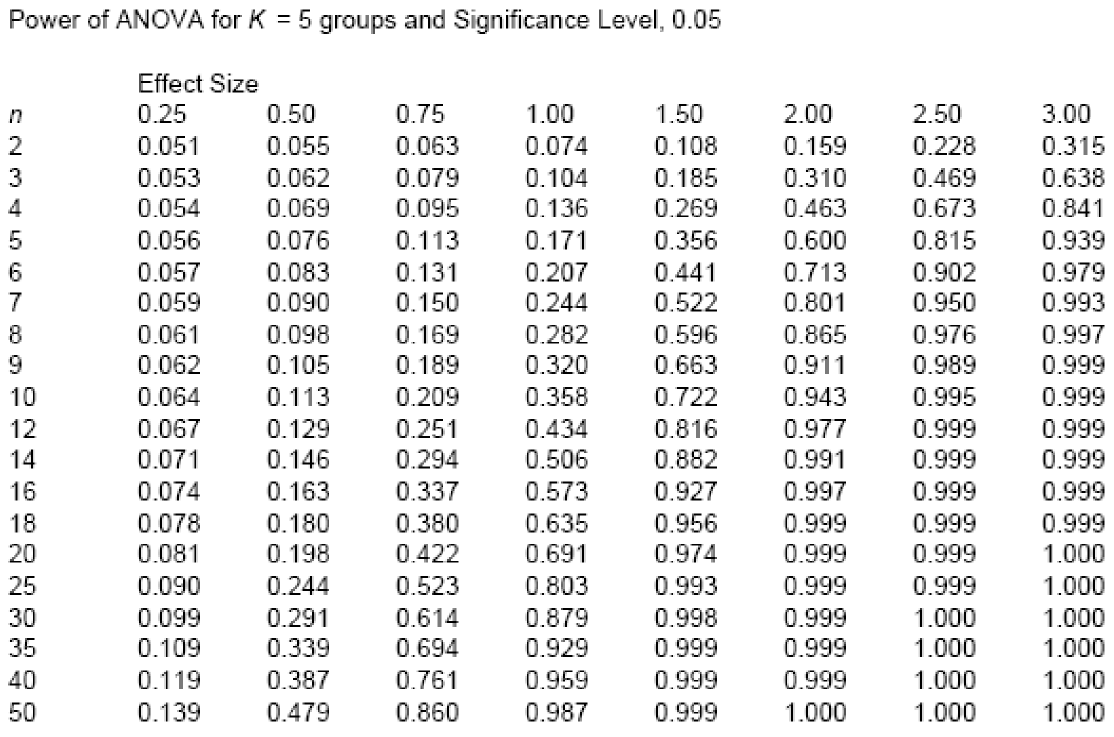
Power of ANOVA for *K* = 4 groups and Significance Level, 0.1

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.103	0.112	0.127	0.148	0.210	0.295	0.398	0.512
3	0.106	0.124	0.155	0.200	0.328	0.495	0.668	0.813
4	0.109	0.136	0.183	0.251	0.439	0.655	0.833	0.938
5	0.111	0.148	0.212	0.302	0.539	0.773	0.922	0.981
6	0.114	0.160	0.240	0.351	0.627	0.856	0.965	0.995
7	0.117	0.173	0.268	0.400	0.702	0.911	0.985	0.998
8	0.120	0.185	0.296	0.447	0.764	0.946	0.994	0.999
9	0.123	0.197	0.324	0.491	0.816	0.968	0.997	0.999
10	0.126	0.210	0.352	0.534	0.857	0.981	0.999	0.999
12	0.132	0.235	0.406	0.613	0.917	0.994	0.999	0.999
14	0.138	0.260	0.458	0.681	0.953	0.998	0.999	0.999
16	0.144	0.285	0.508	0.740	0.974	0.999	0.999	0.999
18	0.150	0.309	0.555	0.790	0.986	0.999	0.999	1.000
20	0.156	0.334	0.598	0.832	0.992	0.999	0.999	1.000
25	0.171	0.395	0.695	0.907	0.998	0.999	1.000	1.000
30	0.187	0.453	0.772	0.950	0.999	0.999	1.000	1.000
35	0.202	0.508	0.833	0.974	0.999	0.999	1.000	1.000
40	0.217	0.560	0.879	0.987	0.999	1.000	1.000	1.000
50	0.249	0.652	0.939	0.996	0.999	1.000	1.000	1.000



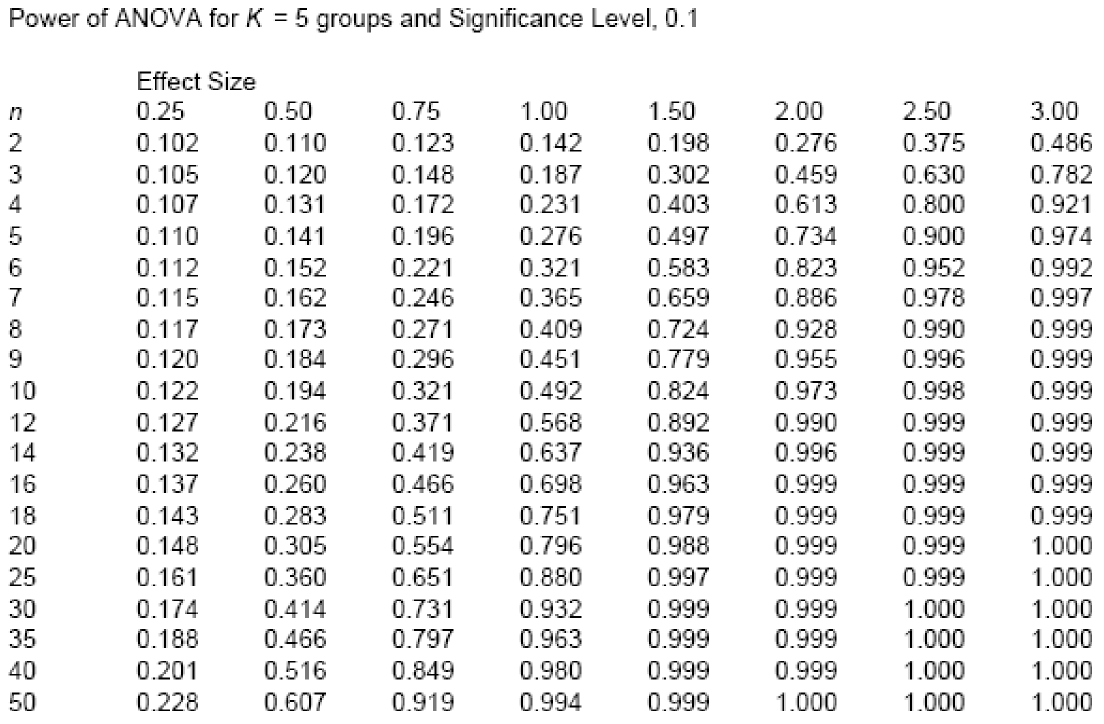
Power of ANOVA for *K* = 4 groups and Significance Level, 0.2

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.204	0.219	0.244	0.278	0.369	0.483	0.606	0.722
3	0.209	0.237	0.283	0.344	0.503	0.677	0.823	0.920
4	0.213	0.253	0.319	0.405	0.613	0.803	0.925	0.979
5	0.217	0.270	0.354	0.462	0.701	0.883	0.969	0.994
6	0.221	0.286	0.387	0.515	0.772	0.932	0.988	0.998
7	0.225	0.301	0.420	0.564	0.828	0.962	0.995	0.999
8	0.229	0.317	0.451	0.609	0.872	0.978	0.998	0.999
9	0.234	0.332	0.482	0.650	0.905	0.988	0.999	0.999
10	0.238	0.348	0.511	0.688	0.930	0.993	0.999	0.999
12	0.246	0.378	0.566	0.754	0.963	0.998	0.999	0.999
14	0.254	0.407	0.616	0.807	0.981	0.999	0.999	0.999
16	0.262	0.435	0.661	0.850	0.990	0.999	0.999	1.000
18	0.270	0.462	0.703	0.885	0.995	0.999	0.999	1.000
20	0.278	0.489	0.739	0.912	0.997	0.999	0.999	1.000
25	0.297	0.551	0.815	0.956	0.999	0.999	1.000	1.000
30	0.317	0.608	0.871	0.978	0.999	0.999	1.000	1.000
35	0.336	0.659	0.911	0.990	0.999	1.000	1.000	1.000
40	0.355	0.705	0.940	0.995	0.999	1.000	1.000	1.000
50	0.392	0.781	0.973	0.999	0.999	1.000	1.000	1.000



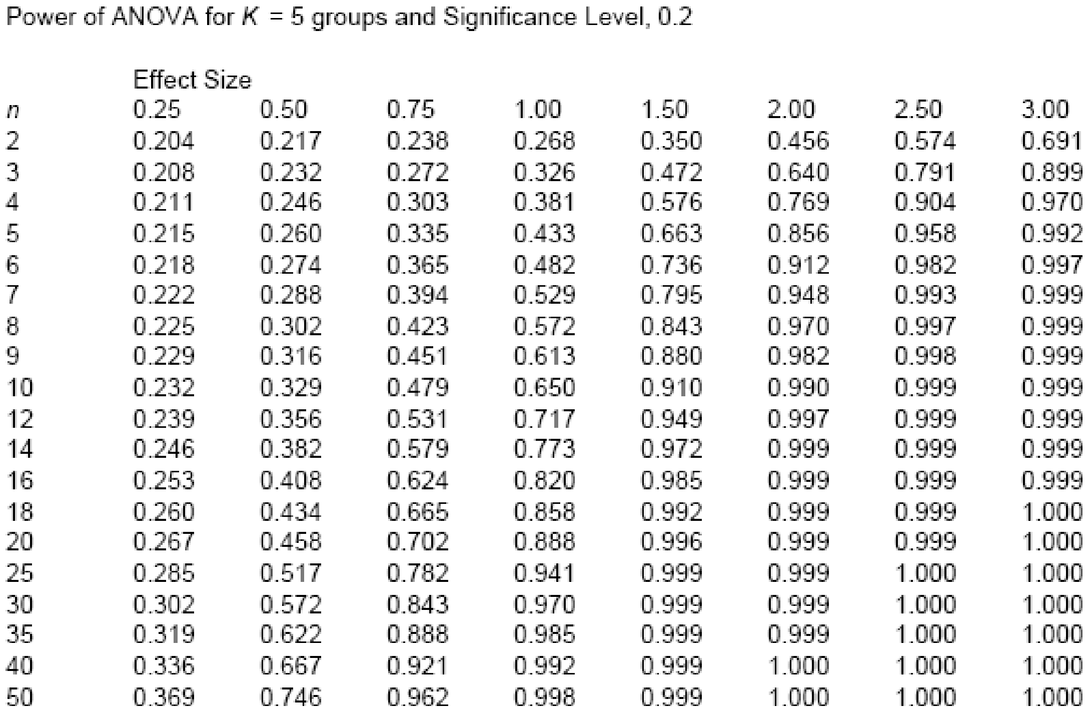
Power of ANOVA for *K* = 5 groups and Significance Level, 0.05

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.051	0.055	0.063	0.074	0.108	0.159	0.228	0.315
3	0.053	0.062	0.079	0.104	0.185	0.310	0.469	0.638
4	0.054	0.069	0.095	0.136	0.269	0.463	0.673	0.841
5	0.056	0.076	0.113	0.171	0.356	0.600	0.815	0.939
6	0.057	0.083	0.131	0.207	0.441	0.713	0.902	0.979
7	0.059	0.090	0.150	0.244	0.522	0.801	0.950	0.993
8	0.061	0.098	0.169	0.282	0.596	0.865	0.976	0.997
9	0.062	0.105	0.189	0.320	0.663	0.911	0.989	0.999
10	0.064	0.113	0.209	0.358	0.722	0.943	0.995	0.999
12	0.067	0.129	0.251	0.434	0.816	0.977	0.999	0.999
14	0.071	0.146	0.294	0.506	0.882	0.991	0.999	0.999
16	0.074	0.163	0.337	0.573	0.927	0.997	0.999	0.999
18	0.078	0.180	0.380	0.635	0.956	0.999	0.999	0.999
20	0.081	0.198	0.422	0.691	0.974	0.999	0.999	1.000
25	0.090	0.244	0.523	0.803	0.993	0.999	0.999	1.000
30	0.099	0.291	0.614	0.879	0.998	0.999	1.000	1.000
35	0.109	0.339	0.694	0.929	0.999	0.999	1.000	1.000
40	0.119	0.387	0.761	0.959	0.999	0.999	1.000	1.000
50	0.139	0.479	0.860	0.987	0.999	1.000	1.000	1.000



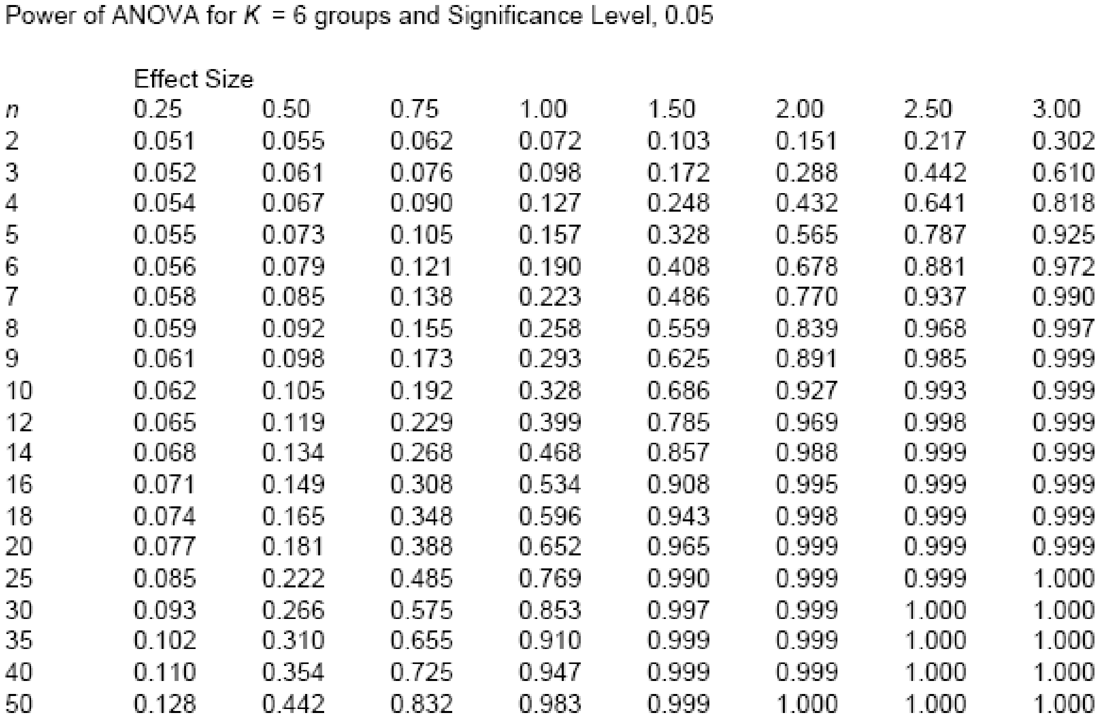
Power of ANOVA for *K* = 5 groups and Significance Level, 0.1

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.102	0.110	0.123	0.142	0.198	0.276	0.375	0.486
3	0.105	0.120	0.148	0.187	0.302	0.459	0.630	0.782
4	0.107	0.131	0.172	0.231	0.403	0.613	0.800	0.921
5	0.110	0.141	0.196	0.276	0.497	0.734	0.900	0.974
6	0.112	0.152	0.221	0.321	0.583	0.823	0.952	0.992
7	0.115	0.162	0.246	0.365	0.659	0.886	0.978	0.997
8	0.117	0.173	0.271	0.409	0.724	0.928	0.990	0.999
9	0.120	0.184	0.296	0.451	0.779	0.955	0.996	0.999
10	0.122	0.194	0.321	0.492	0.824	0.973	0.998	0.999
12	0.127	0.216	0.371	0.568	0.892	0.990	0.999	0.999
14	0.132	0.238	0.419	0.637	0.936	0.996	0.999	0.999
16	0.137	0.260	0.466	0.698	0.963	0.999	0.999	0.999
18	0.143	0.283	0.511	0.751	0.979	0.999	0.999	0.999
20	0.148	0.305	0.554	0.796	0.988	0.999	0.999	1.000
25	0.161	0.360	0.651	0.880	0.997	0.999	0.999	1.000
30	0.174	0.414	0.731	0.932	0.999	0.999	1.000	1.000
35	0.188	0.466	0.797	0.963	0.999	0.999	1.000	1.000
40	0.201	0.516	0.849	0.980	0.999	0.999	1.000	1.000
50	0.228	0.607	0.919	0.994	0.999	1.000	1.000	1.000



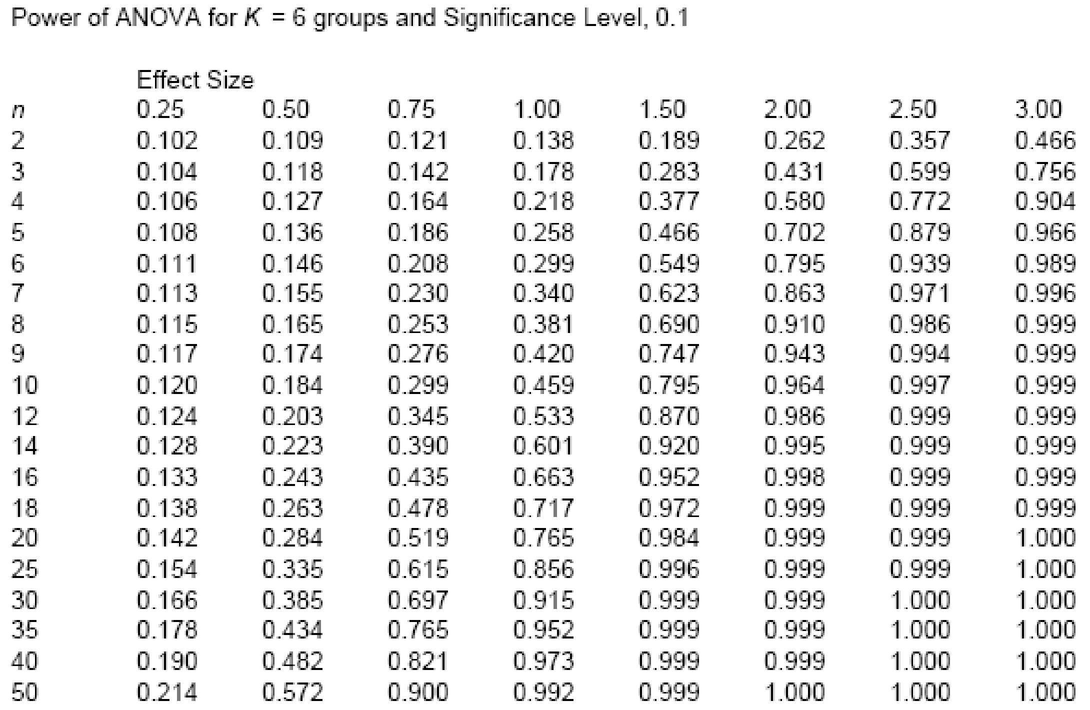
Power of ANOVA for *K* = 5 groups and Significance Level, 0.2

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.204	0.217	0.238	0.268	0.350	0.456	0.574	0.691
3	0.208	0.232	0.272	0.326	0.472	0.640	0.791	0.899
4	0.211	0.246	0.303	0.381	0.576	0.769	0.904	0.970
5	0.215	0.260	0.335	0.433	0.663	0.856	0.958	0.992
6	0.218	0.274	0.365	0.482	0.736	0.912	0.982	0.997
7	0.222	0.288	0.394	0.529	0.795	0.948	0.993	0.999
8	0.225	0.302	0.423	0.572	0.843	0.970	0.997	0.999
9	0.229	0.316	0.451	0.613	0.880	0.982	0.998	0.999
10	0.232	0.329	0.479	0.650	0.910	0.990	0.999	0.999
12	0.239	0.356	0.531	0.717	0.949	0.997	0.999	0.999
14	0.246	0.382	0.579	0.773	0.972	0.999	0.999	0.999
16	0.253	0.408	0.624	0.820	0.985	0.999	0.999	0.999
18	0.260	0.434	0.665	0.858	0.992	0.999	0.999	1.000
20	0.267	0.458	0.702	0.888	0.996	0.999	0.999	1.000
25	0.285	0.517	0.782	0.941	0.999	0.999	1.000	1.000
30	0.302	0.572	0.843	0.970	0.999	0.999	1.000	1.000
35	0.319	0.622	0.888	0.985	0.999	0.999	1.000	1.000
40	0.336	0.667	0.921	0.992	0.999	1.000	1.000	1.000
50	0.369	0.746	0.962	0.998	0.999	1.000	1.000	1.000



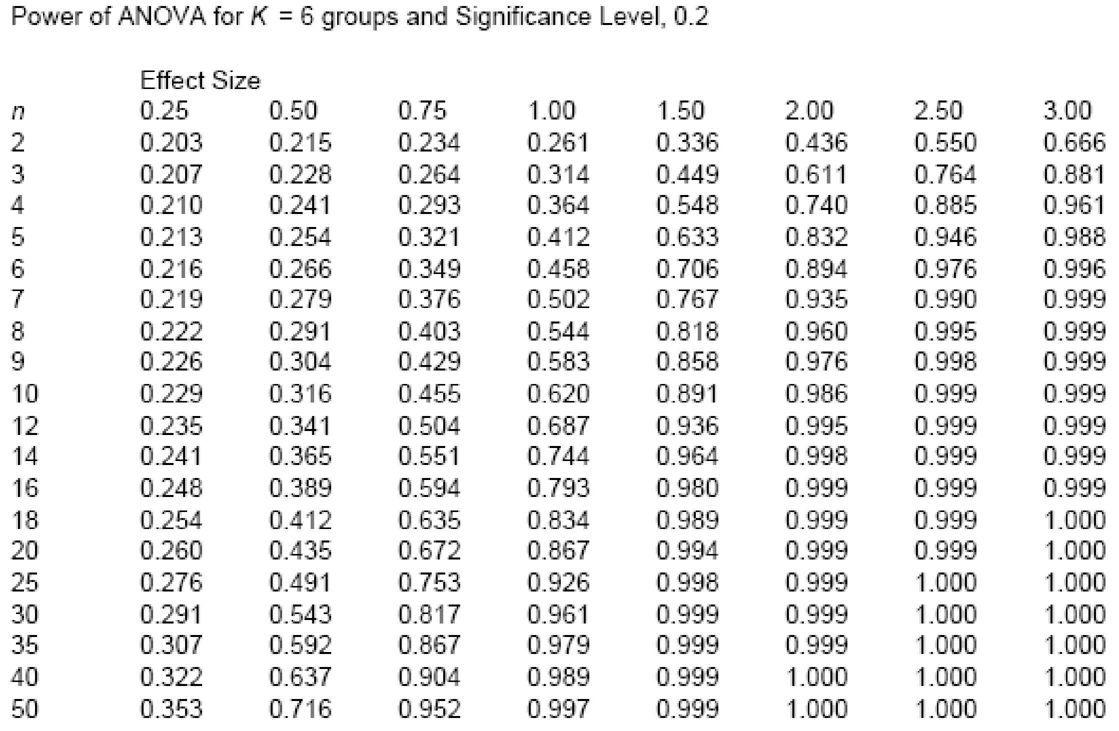
Power of ANOVA for *K* = 6 groups and Significance Level, 0.05

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.051	0.055	0.062	0.072	0.103	0.151	0.217	0.302
3	0.052	0.061	0.076	0.098	0.172	0.288	0.442	0.610
4	0.054	0.067	0.090	0.127	0.248	0.432	0.641	0.818
5	0.055	0.073	0.105	0.157	0.328	0.565	0.787	0.925
6	0.056	0.079	0.121	0.190	0.408	0.678	0.881	0.972
7	0.058	0.085	0.138	0.223	0.486	0.770	0.937	0.990
8	0.059	0.092	0.155	0.258	0.559	0.839	0.968	0.997
9	0.061	0.098	0.173	0.293	0.625	0.891	0.985	0.999
10	0.062	0.105	0.192	0.328	0.686	0.927	0.993	0.999
12	0.065	0.119	0.229	0.399	0.785	0.969	0.998	0.999
14	0.068	0.134	0.268	0.468	0.857	0.988	0.999	0.999
16	0.071	0.149	0.308	0.534	0.908	0.995	0.999	0.999
18	0.074	0.165	0.348	0.596	0.943	0.998	0.999	0.999
20	0.077	0.181	0.388	0.652	0.965	0.999	0.999	0.999
25	0.085	0.222	0.485	0.769	0.990	0.999	0.999	1.000
30	0.093	0.266	0.575	0.853	0.997	0.999	1.000	1.000
35	0.102	0.310	0.655	0.910	0.999	0.999	1.000	1.000
40	0.110	0.354	0.725	0.947	0.999	0.999	1.000	1.000
50	0.128	0.442	0.832	0.983	0.999	1.000	1.000	1.000



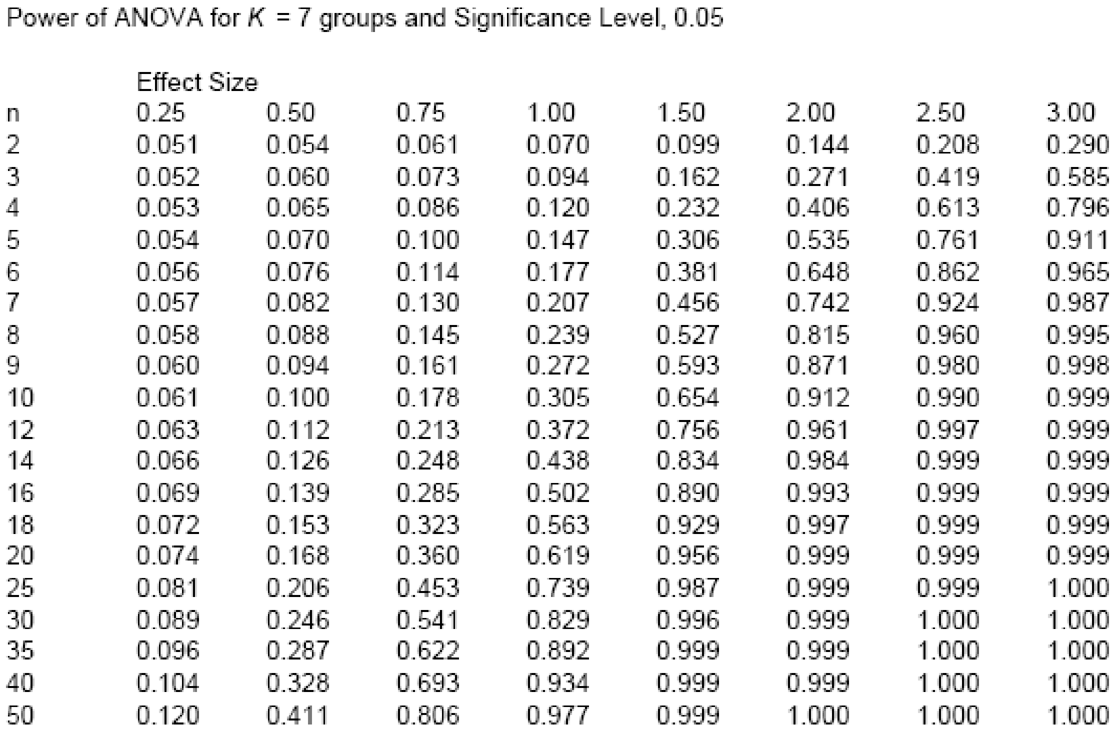
Power of ANOVA for *K* = 6 groups and Significance Level, 0.1

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.102	0.109	0.121	0.138	0.189	0.262	0.357	0.466
3	0.104	0.118	0.142	0.178	0.283	0.431	0.599	0.756
4	0.106	0.127	0.164	0.218	0.377	0.580	0.772	0.904
5	0.108	0.136	0.186	0.258	0.466	0.702	0.879	0.966
6	0.111	0.146	0.208	0.299	0.549	0.795	0.939	0.989
7	0.113	0.155	0.230	0.340	0.623	0.863	0.971	0.996
8	0.115	0.165	0.253	0.381	0.690	0.910	0.986	0.999
9	0.117	0.174	0.276	0.420	0.747	0.943	0.994	0.999
10	0.120	0.184	0.299	0.459	0.795	0.964	0.997	0.999
12	0.124	0.203	0.345	0.533	0.870	0.986	0.999	0.999
14	0.128	0.223	0.390	0.601	0.920	0.995	0.999	0.999
16	0.133	0.243	0.435	0.663	0.952	0.998	0.999	0.999
18	0.138	0.263	0.478	0.717	0.972	0.999	0.999	0.999
20	0.142	0.284	0.519	0.765	0.984	0.999	0.999	1.000
25	0.154	0.335	0.615	0.856	0.996	0.999	0.999	1.000
30	0.166	0.385	0.697	0.915	0.999	0.999	1.000	1.000
35	0.178	0.434	0.765	0.952	0.999	0.999	1.000	1.000
40	0.190	0.482	0.821	0.973	0.999	0.999	1.000	1.000
50	0.214	0.572	0.900	0.992	0.999	1.000	1.000	1.000



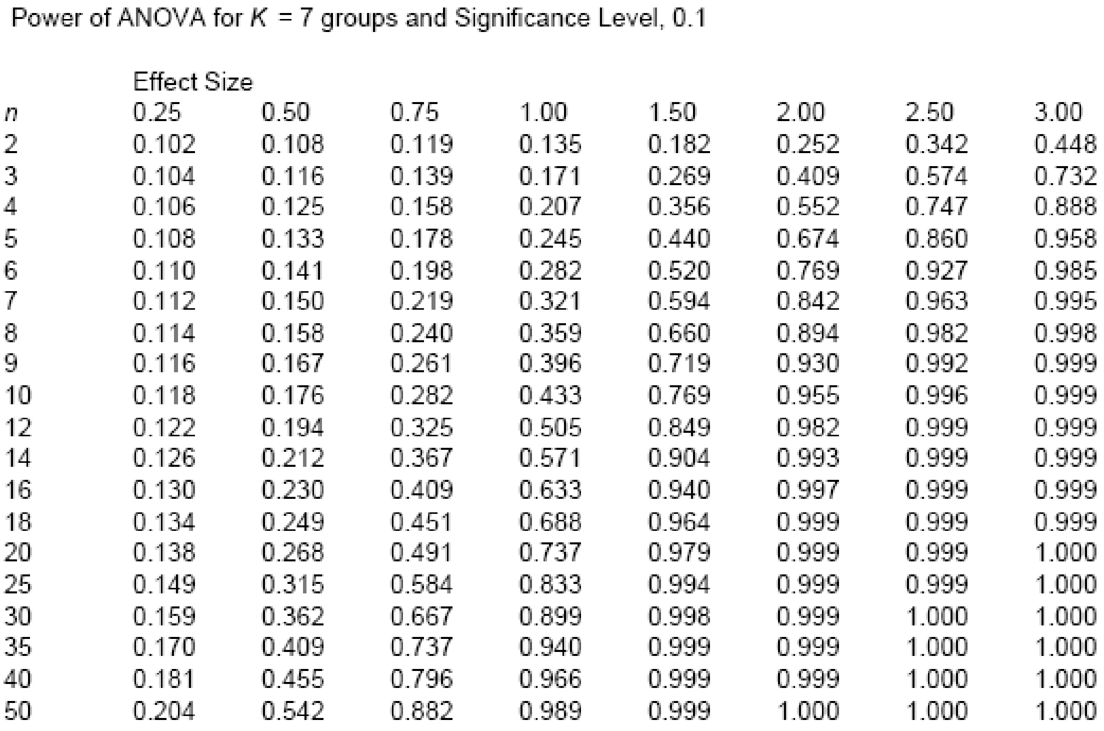
Power of ANOVA for *K* = 6 groups and Significance Level, 0.2

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.203	0.215	0.234	0.261	0.336	0.436	0.550	0.666
3	0.207	0.228	0.264	0.314	0.449	0.611	0.764	0.881
4	0.210	0.241	0.293	0.364	0.548	0.740	0.885	0.961
5	0.213	0.254	0.321	0.412	0.633	0.832	0.946	0.988
6	0.216	0.266	0.349	0.458	0.706	0.894	0.976	0.996
7	0.219	0.279	0.376	0.502	0.767	0.935	0.990	0.999
8	0.222	0.291	0.403	0.544	0.818	0.960	0.995	0.999
9	0.226	0.304	0.429	0.583	0.858	0.976	0.998	0.999
10	0.229	0.316	0.455	0.620	0.891	0.986	0.999	0.999
12	0.235	0.341	0.504	0.687	0.936	0.995	0.999	0.999
14	0.241	0.365	0.551	0.744	0.964	0.998	0.999	0.999
16	0.248	0.389	0.594	0.793	0.980	0.999	0.999	0.999
18	0.254	0.412	0.635	0.834	0.989	0.999	0.999	1.000
20	0.260	0.435	0.672	0.867	0.994	0.999	0.999	1.000
25	0.276	0.491	0.753	0.926	0.998	0.999	1.000	1.000
30	0.291	0.543	0.817	0.961	0.999	0.999	1.000	1.000
35	0.307	0.592	0.867	0.979	0.999	0.999	1.000	1.000
40	0.322	0.637	0.904	0.989	0.999	1.000	1.000	1.000
50	0.353	0.716	0.952	0.997	0.999	1.000	1.000	1.000



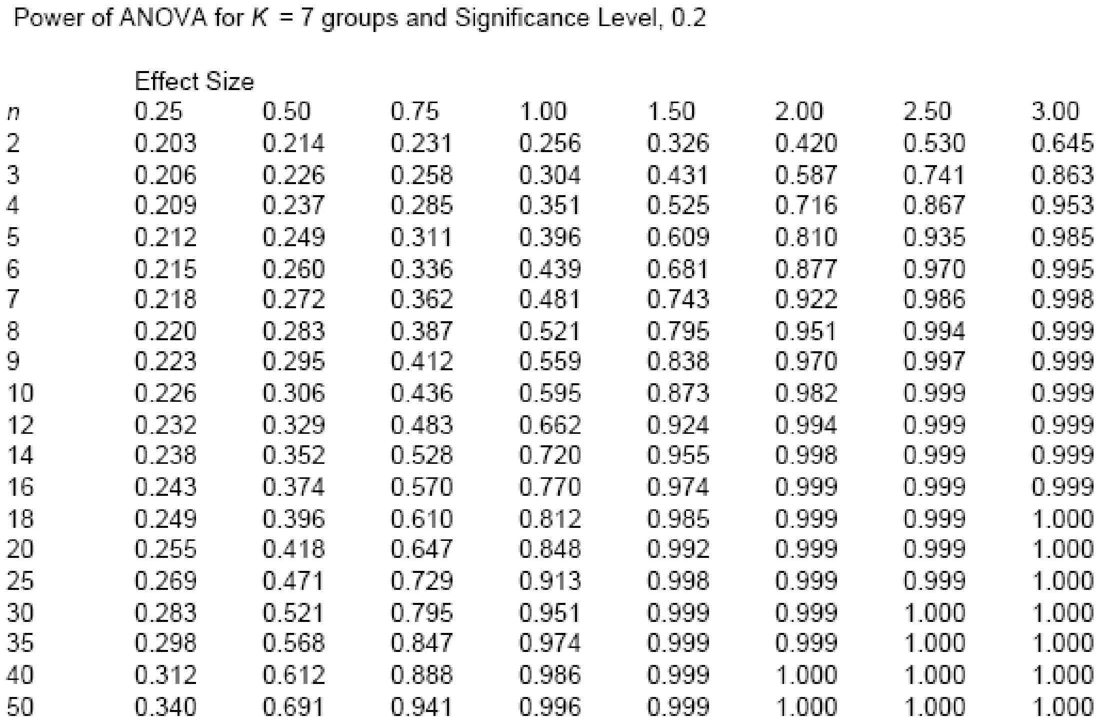
Power of ANOVA for *K* = 7 groups and Significance Level, 0.05

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.051	0.054	0.061	0.070	0.099	0.144	0.208	0.290
3	0.052	0.060	0.073	0.094	0.162	0.271	0.419	0.585
4	0.053	0.065	0.086	0.120	0.232	0.406	0.613	0.796
5	0.054	0.070	0.100	0.147	0.306	0.535	0.761	0.911
6	0.056	0.076	0.114	0.177	0.381	0.648	0.862	0.965
7	0.057	0.082	0.130	0.207	0.456	0.742	0.924	0.987
8	0.058	0.088	0.145	0.239	0.527	0.815	0.960	0.995
9	0.060	0.094	0.161	0.272	0.593	0.871	0.980	0.998
10	0.061	0.100	0.178	0.305	0.654	0.912	0.990	0.999
12	0.063	0.112	0.213	0.372	0.756	0.961	0.997	0.999
14	0.066	0.126	0.248	0.438	0.834	0.984	0.999	0.999
16	0.069	0.139	0.285	0.502	0.890	0.993	0.999	0.999
18	0.072	0.153	0.323	0.563	0.929	0.997	0.999	0.999
20	0.074	0.168	0.360	0.619	0.956	0.999	0.999	0.999
25	0.081	0.206	0.453	0.739	0.987	0.999	0.999	1.000
30	0.089	0.246	0.541	0.829	0.996	0.999	1.000	1.000
35	0.096	0.287	0.622	0.892	0.999	0.999	1.000	1.000
40	0.104	0.328	0.693	0.934	0.999	0.999	1.000	1.000
50	0.120	0.411	0.806	0.977	0.999	1.000	1.000	1.000



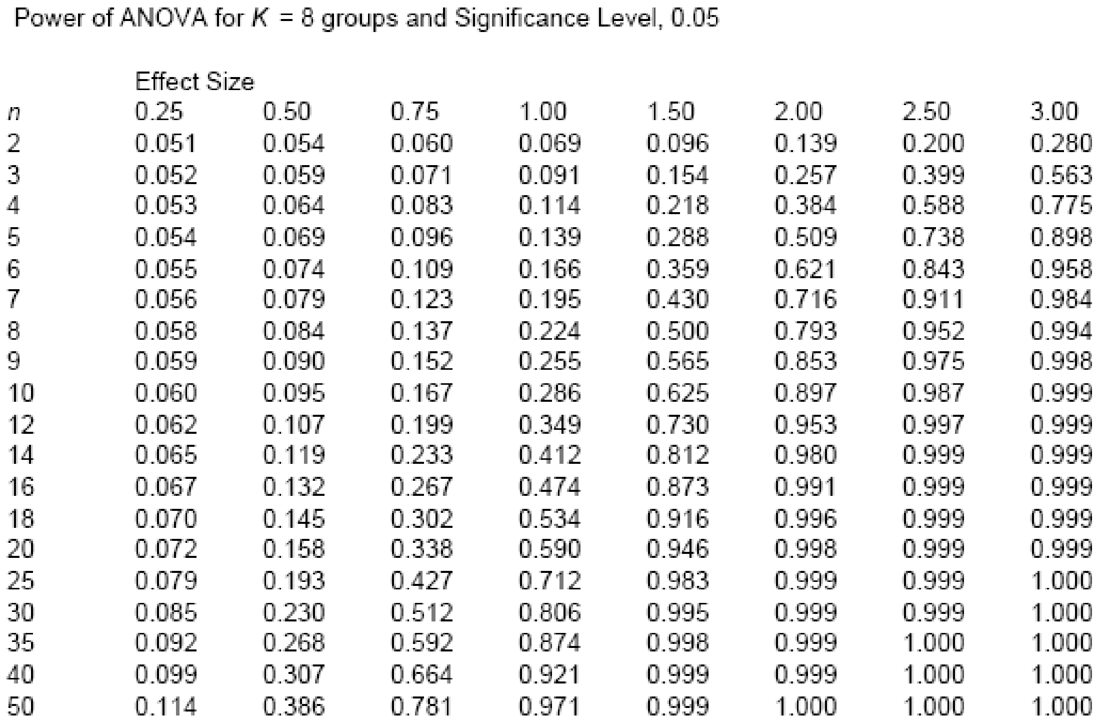
Power of ANOVA for *K* = 7 groups and Significance Level, 0.1

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.102	0.108	0.119	0.135	0.182	0.252	0.342	0.448
3	0.104	0.116	0.139	0.171	0.269	0.409	0.574	0.732
4	0.106	0.125	0.158	0.207	0.356	0.552	0.747	0.888
5	0.108	0.133	0.178	0.245	0.440	0.674	0.860	0.958
6	0.110	0.141	0.198	0.282	0.520	0.769	0.927	0.985
7	0.112	0.150	0.219	0.321	0.594	0.842	0.963	0.995
8	0.114	0.158	0.240	0.359	0.660	0.894	0.982	0.998
9	0.116	0.167	0.261	0.396	0.719	0.930	0.992	0.999
10	0.118	0.176	0.282	0.433	0.769	0.955	0.996	0.999
12	0.122	0.194	0.325	0.505	0.849	0.982	0.999	0.999
14	0.126	0.212	0.367	0.571	0.904	0.993	0.999	0.999
16	0.130	0.230	0.409	0.633	0.940	0.997	0.999	0.999
18	0.134	0.249	0.451	0.688	0.964	0.999	0.999	0.999
20	0.138	0.268	0.491	0.737	0.979	0.999	0.999	1.000
25	0.149	0.315	0.584	0.833	0.994	0.999	0.999	1.000
30	0.159	0.362	0.667	0.899	0.998	0.999	1.000	1.000
35	0.170	0.409	0.737	0.940	0.999	0.999	1.000	1.000
40	0.181	0.455	0.796	0.966	0.999	0.999	1.000	1.000
50	0.204	0.542	0.882	0.989	0.999	1.000	1.000	1.000



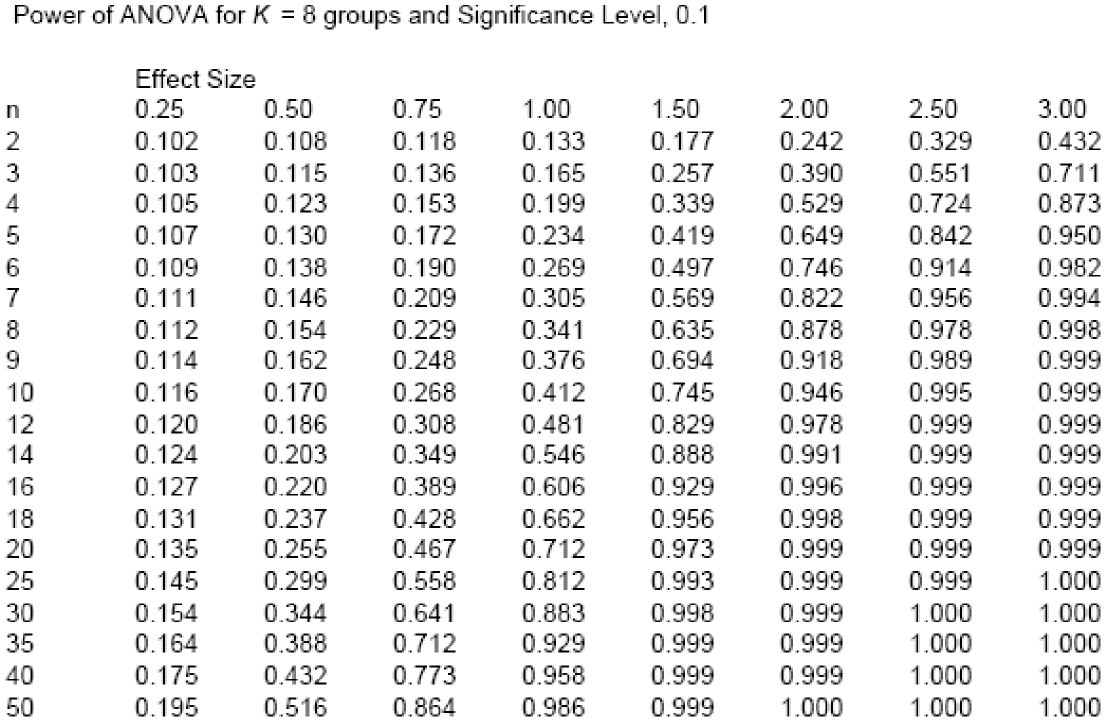
Power of ANOVA for *K* = 7 groups and Significance Level, 0.2

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.203	0.214	0.231	0.256	0.326	0.420	0.530	0.645
3	0.206	0.226	0.258	0.304	0.431	0.587	0.741	0.863
4	0.209	0.237	0.285	0.351	0.525	0.716	0.867	0.953
5	0.212	0.249	0.311	0.396	0.609	0.810	0.935	0.985
6	0.215	0.260	0.336	0.439	0.681	0.877	0.970	0.995
7	0.218	0.272	0.362	0.481	0.743	0.922	0.986	0.998
8	0.220	0.283	0.387	0.521	0.795	0.951	0.994	0.999
9	0.223	0.295	0.412	0.559	0.838	0.970	0.997	0.999
10	0.226	0.306	0.436	0.595	0.873	0.982	0.999	0.999
12	0.232	0.329	0.483	0.662	0.924	0.994	0.999	0.999
14	0.238	0.352	0.528	0.720	0.955	0.998	0.999	0.999
16	0.243	0.374	0.570	0.770	0.974	0.999	0.999	0.999
18	0.249	0.396	0.610	0.812	0.985	0.999	0.999	1.000
20	0.255	0.418	0.647	0.848	0.992	0.999	0.999	1.000
25	0.269	0.471	0.729	0.913	0.998	0.999	0.999	1.000
30	0.283	0.521	0.795	0.951	0.999	0.999	1.000	1.000
35	0.298	0.568	0.847	0.974	0.999	0.999	1.000	1.000
40	0.312	0.612	0.888	0.986	0.999	1.000	1.000	1.000
50	0.340	0.691	0.941	0.996	0.999	1.000	1.000	1.000



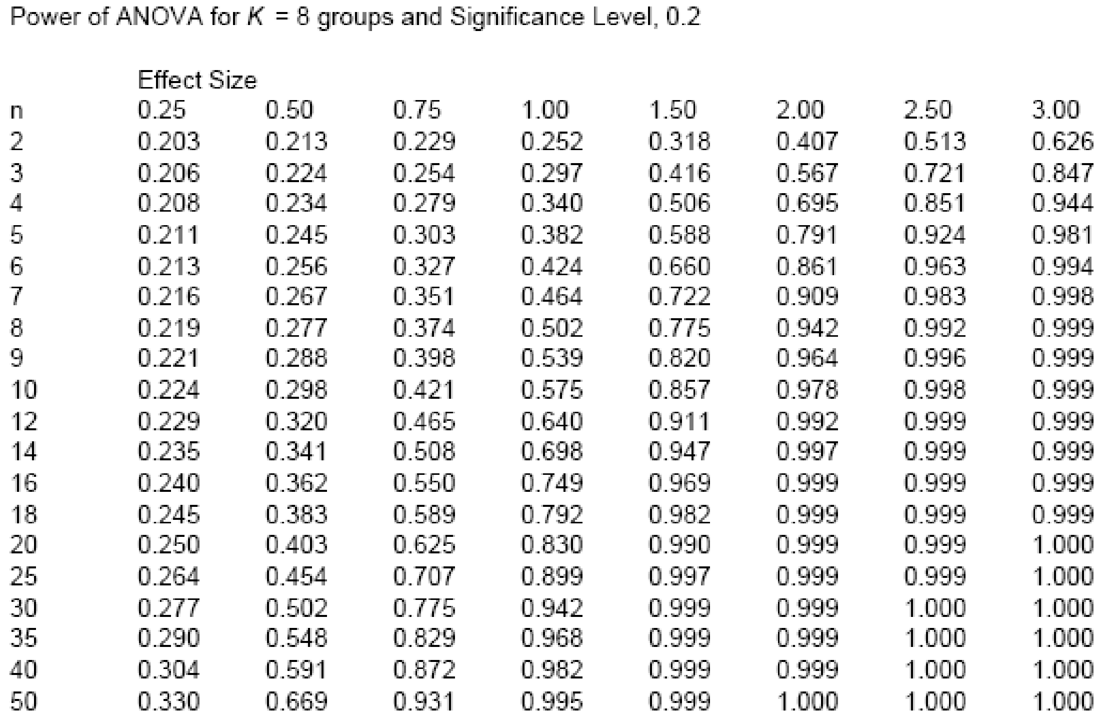
Power of ANOVA for *K* = 8 groups and Significance Level, 0.05

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.051	0.054	0.060	0.069	0.096	0.139	0.200	0.280
3	0.052	0.059	0.071	0.091	0.154	0.257	0.399	0.563
4	0.053	0.064	0.083	0.114	0.218	0.384	0.588	0.775
5	0.054	0.069	0.096	0.139	0.288	0.509	0.738	0.898
6	0.055	0.074	0.109	0.166	0.359	0.621	0.843	0.958
7	0.056	0.079	0.123	0.195	0.430	0.716	0.911	0.984
8	0.058	0.084	0.137	0.224	0.500	0.793	0.952	0.994
9	0.059	0.090	0.152	0.255	0.565	0.853	0.975	0.998
10	0.060	0.095	0.167	0.286	0.625	0.897	0.987	0.999
12	0.062	0.107	0.199	0.349	0.730	0.953	0.997	0.999
14	0.065	0.119	0.233	0.412	0.812	0.980	0.999	0.999
16	0.067	0.132	0.267	0.474	0.873	0.991	0.999	0.999
18	0.070	0.145	0.302	0.534	0.916	0.996	0.999	0.999
20	0.072	0.158	0.338	0.590	0.946	0.998	0.999	0.999
25	0.079	0.193	0.427	0.712	0.983	0.999	0.999	1.000
30	0.085	0.230	0.512	0.806	0.995	0.999	0.999	1.000
35	0.092	0.268	0.592	0.874	0.998	0.999	1.000	1.000
40	0.099	0.307	0.664	0.921	0.999	0.999	1.000	1.000
50	0.114	0.386	0.781	0.971	0.999	1.000	1.000	1.000



Power of ANOVA for *K* = 8 groups and Significance Level, 0.1

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.102	0.108	0.118	0.133	0.177	0.242	0.329	0.432
3	0.103	0.115	0.136	0.165	0.257	0.390	0.551	0.711
4	0.105	0.123	0.153	0.199	0.339	0.529	0.724	0.873
5	0.107	0.130	0.172	0.234	0.419	0.649	0.842	0.950
6	0.109	0.138	0.190	0.269	0.497	0.746	0.914	0.982
7	0.111	0.146	0.209	0.305	0.569	0.822	0.956	0.994
8	0.112	0.154	0.229	0.341	0.635	0.878	0.978	0.998
9	0.114	0.162	0.248	0.376	0.694	0.918	0.989	0.999
10	0.116	0.170	0.268	0.412	0.745	0.946	0.995	0.999
12	0.120	0.186	0.308	0.481	0.829	0.978	0.999	0.999
14	0.124	0.203	0.349	0.546	0.888	0.991	0.999	0.999
16	0.127	0.220	0.389	0.606	0.929	0.996	0.999	0.999
18	0.131	0.237	0.428	0.662	0.956	0.998	0.999	0.999
20	0.135	0.255	0.467	0.712	0.973	0.999	0.999	0.999
25	0.145	0.299	0.558	0.812	0.993	0.999	0.999	1.000
30	0.154	0.344	0.641	0.883	0.998	0.999	1.000	1.000
35	0.164	0.388	0.712	0.929	0.999	0.999	1.000	1.000
40	0.175	0.432	0.773	0.958	0.999	0.999	1.000	1.000
50	0.195	0.516	0.864	0.986	0.999	1.000	1.000	1.000



Power of ANOVA for *K* = 8 groups and Significance Level, 0.2

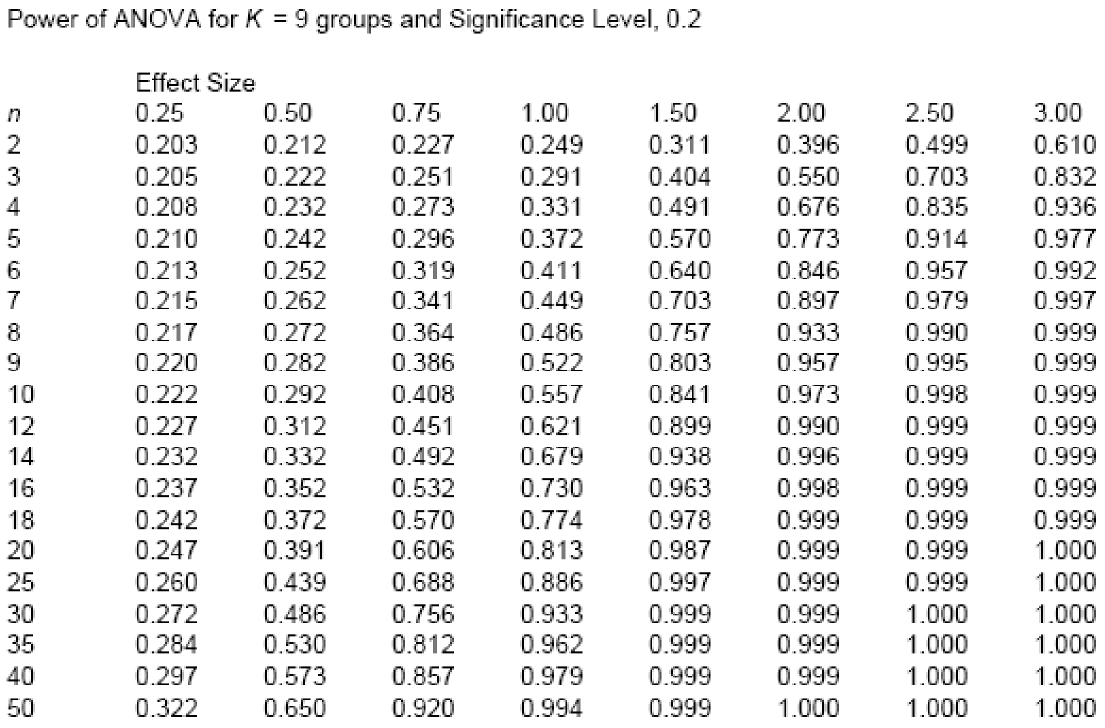
	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.203	0.213	0.229	0.252	0.318	0.407	0.513	0.626
3	0.206	0.224	0.254	0.297	0.416	0.567	0.721	0.847
4	0.208	0.234	0.279	0.340	0.506	0.695	0.851	0.944
5	0.211	0.245	0.303	0.382	0.588	0.791	0.924	0.981
6	0.213	0.256	0.327	0.424	0.660	0.861	0.963	0.994
7	0.216	0.267	0.351	0.464	0.722	0.909	0.983	0.998
8	0.219	0.277	0.374	0.502	0.775	0.942	0.992	0.999
9	0.221	0.288	0.398	0.539	0.820	0.964	0.996	0.999
10	0.224	0.298	0.421	0.575	0.857	0.978	0.998	0.999
12	0.229	0.320	0.465	0.640	0.911	0.992	0.999	0.999
14	0.235	0.341	0.508	0.698	0.947	0.997	0.999	0.999
16	0.240	0.362	0.550	0.749	0.969	0.999	0.999	0.999
18	0.245	0.383	0.589	0.792	0.982	0.999	0.999	0.999
20	0.250	0.403	0.625	0.830	0.990	0.999	0.999	1.000
25	0.264	0.454	0.707	0.899	0.997	0.999	0.999	1.000
30	0.277	0.502	0.775	0.942	0.999	0.999	1.000	1.000
35	0.290	0.548	0.829	0.968	0.999	0.999	1.000	1.000
40	0.304	0.591	0.872	0.982	0.999	0.999	1.000	1.000
50	0.330	0.669	0.931	0.995	0.999	1.000	1.000	1.000

Power of ANOVA for *K* = 9 groups and Significance Level, 0.05

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.051	0.054	0.060	0.068	0.094	0.134	0.193	0.270
3	0.052	0.058	0.070	0.088	0.147	0.244	0.381	0.544
4	0.053	0.063	0.081	0.110	0.207	0.366	0.565	0.756
5	0.054	0.067	0.093	0.133	0.272	0.486	0.716	0.884
6	0.055	0.072	0.105	0.158	0.340	0.597	0.826	0.950
7	0.056	0.077	0.117	0.184	0.409	0.693	0.898	0.980
8	0.057	0.082	0.131	0.212	0.476	0.772	0.943	0.992
9	0.058	0.087	0.144	0.240	0.540	0.835	0.970	0.997
10	0.059	0.092	0.159	0.270	0.600	0.883	0.984	0.999
12	0.061	0.103	0.188	0.330	0.706	0.944	0.996	0.999
14	0.063	0.114	0.219	0.390	0.791	0.975	0.999	0.999
16	0.066	0.125	0.252	0.450	0.855	0.989	0.999	0.999
18	0.068	0.137	0.285	0.509	0.903	0.995	0.999	0.999
20	0.070	0.150	0.319	0.564	0.936	0.998	0.999	0.999
25	0.076	0.182	0.404	0.687	0.979	0.999	0.999	1.000
30	0.082	0.217	0.487	0.784	0.994	0.999	0.999	1.000
35	0.089	0.253	0.566	0.857	0.998	0.999	1.000	1.000
40	0.095	0.289	0.638	0.908	0.999	0.999	1.000	1.000
50	0.109	0.365	0.758	0.965	0.999	1.000	1.000	1.000

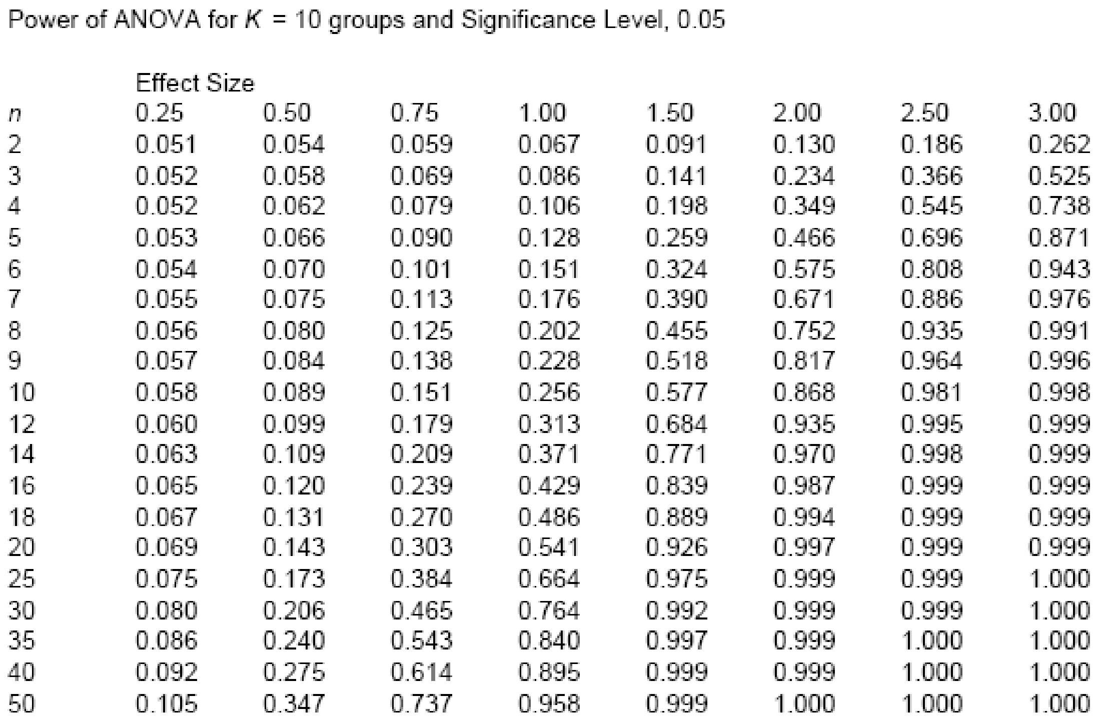
Power of ANOVA for *K* = 9 groups and Significance Level, 0.1

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.101	0.107	0.117	0.131	0.173	0.235	0.318	0.419
3	0.103	0.114	0.133	0.161	0.247	0.375	0.532	0.692
4	0.105	0.121	0.150	0.192	0.324	0.508	0.704	0.858
5	0.106	0.128	0.167	0.225	0.401	0.627	0.825	0.941
6	0.108	0.135	0.184	0.258	0.476	0.725	0.902	0.978
7	0.110	0.143	0.202	0.292	0.547	0.803	0.948	0.992
8	0.111	0.150	0.220	0.326	0.612	0.862	0.973	0.997
9	0.113	0.157	0.238	0.360	0.671	0.906	0.987	0.999
10	0.115	0.165	0.257	0.394	0.724	0.937	0.993	0.999
12	0.118	0.180	0.295	0.460	0.810	0.973	0.998	0.999
14	0.122	0.196	0.333	0.524	0.873	0.989	0.999	0.999
16	0.125	0.212	0.371	0.583	0.918	0.995	0.999	0.999
18	0.129	0.228	0.409	0.638	0.948	0.998	0.999	0.999
20	0.132	0.244	0.446	0.689	0.968	0.999	0.999	0.999
25	0.141	0.286	0.536	0.792	0.991	0.999	0.999	1.000
30	0.150	0.328	0.617	0.867	0.997	0.999	1.000	1.000
35	0.160	0.371	0.689	0.917	0.999	0.999	1.000	1.000
40	0.169	0.413	0.751	0.950	0.999	0.999	1.000	1.000
50	0.189	0.495	0.847	0.983	0.999	1.000	1.000	1.000



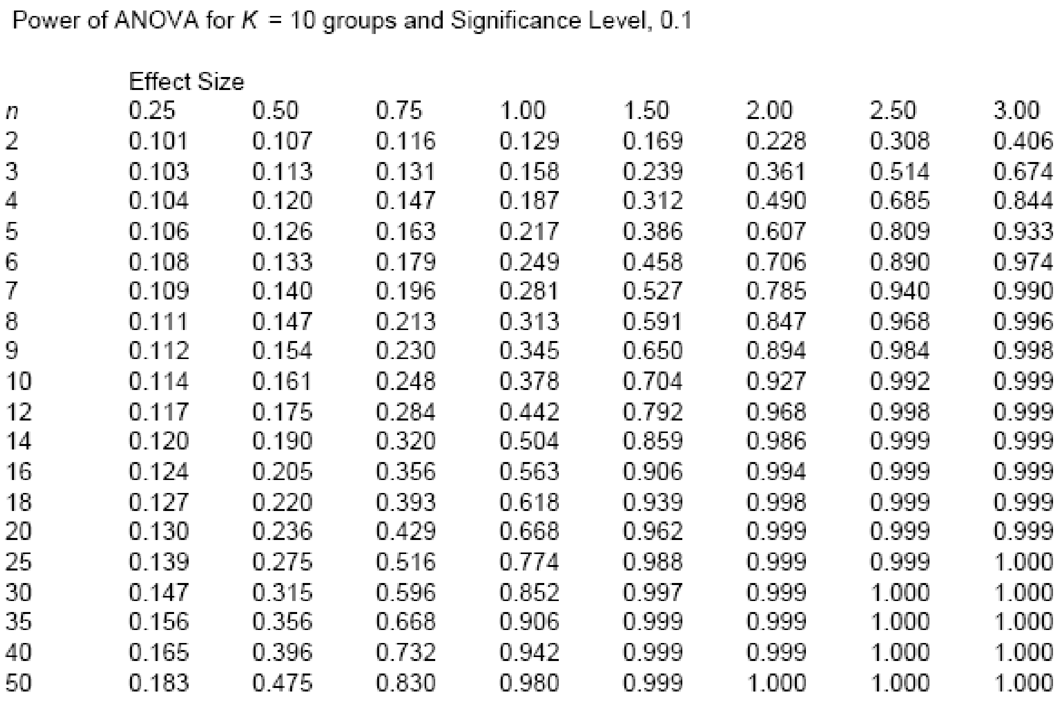
Power of ANOVA for *K* = 9 groups and Significance Level, 0.2

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.203	0.212	0.227	0.249	0.311	0.396	0.499	0.610
3	0.205	0.222	0.251	0.291	0.404	0.550	0.703	0.832
4	0.208	0.232	0.273	0.331	0.491	0.676	0.835	0.936
5	0.210	0.242	0.296	0.372	0.570	0.773	0.914	0.977
6	0.213	0.252	0.319	0.411	0.640	0.846	0.957	0.992
7	0.215	0.262	0.341	0.449	0.703	0.897	0.979	0.997
8	0.217	0.272	0.364	0.486	0.757	0.933	0.990	0.999
9	0.220	0.282	0.386	0.522	0.803	0.957	0.995	0.999
10	0.222	0.292	0.408	0.557	0.841	0.973	0.998	0.999
12	0.227	0.312	0.451	0.621	0.899	0.990	0.999	0.999
14	0.232	0.332	0.492	0.679	0.938	0.996	0.999	0.999
16	0.237	0.352	0.532	0.730	0.963	0.998	0.999	0.999
18	0.242	0.372	0.570	0.774	0.978	0.999	0.999	0.999
20	0.247	0.391	0.606	0.813	0.987	0.999	0.999	1.000
25	0.260	0.439	0.688	0.886	0.997	0.999	0.999	1.000
30	0.272	0.486	0.756	0.933	0.999	0.999	1.000	1.000
35	0.284	0.530	0.812	0.962	0.999	0.999	1.000	1.000
40	0.297	0.573	0.857	0.979	0.999	0.999	1.000	1.000
50	0.322	0.650	0.920	0.994	0.999	1.000	1.000	1.000



Power of ANOVA for *K* = 10 groups and Significance Level, 0.05

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.051	0.054	0.059	0.067	0.091	0.130	0.186	0.262
3	0.052	0.058	0.069	0.086	0.141	0.234	0.366	0.525
4	0.052	0.062	0.079	0.106	0.198	0.349	0.545	0.738
5	0.053	0.066	0.090	0.128	0.259	0.466	0.696	0.871
6	0.054	0.070	0.101	0.151	0.324	0.575	0.808	0.943
7	0.055	0.075	0.113	0.176	0.390	0.671	0.886	0.976
8	0.056	0.080	0.125	0.202	0.455	0.752	0.935	0.991
9	0.057	0.084	0.138	0.228	0.518	0.817	0.964	0.996
10	0.058	0.089	0.151	0.256	0.577	0.868	0.981	0.998
12	0.060	0.099	0.179	0.313	0.684	0.935	0.995	0.999
14	0.063	0.109	0.209	0.371	0.771	0.970	0.998	0.999
16	0.065	0.120	0.239	0.429	0.839	0.987	0.999	0.999
18	0.067	0.131	0.270	0.486	0.889	0.994	0.999	0.999
20	0.069	0.143	0.303	0.541	0.926	0.997	0.999	0.999
25	0.075	0.173	0.384	0.664	0.975	0.999	0.999	1.000
30	0.080	0.206	0.465	0.764	0.992	0.999	0.999	1.000
35	0.086	0.240	0.543	0.840	0.997	0.999	1.000	1.000
40	0.092	0.275	0.614	0.895	0.999	0.999	1.000	1.000
50	0.105	0.347	0.737	0.958	0.999	1.000	1.000	1.000



Power of ANOVA for *K* = 10 groups and Significance Level, 0.1

	Effect Size							
n	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
2	0.101	0.107	0.116	0.129	0.169	0.228	0.308	0.406
3	0.103	0.113	0.131	0.158	0.239	0.361	0.514	0.674
4	0.104	0.120	0.147	0.187	0.312	0.490	0.685	0.844
5	0.106	0.126	0.163	0.217	0.386	0.607	0.809	0.933
6	0.108	0.133	0.179	0.249	0.458	0.706	0.890	0.974
7	0.109	0.140	0.196	0.281	0.527	0.785	0.940	0.990
8	0.111	0.147	0.213	0.313	0.591	0.847	0.968	0.996
9	0.112	0.154	0.230	0.345	0.650	0.894	0.984	0.998
10	0.114	0.161	0.248	0.378	0.704	0.927	0.992	0.999
12	0.117	0.175	0.284	0.442	0.792	0.968	0.998	0.999
14	0.120	0.190	0.320	0.504	0.859	0.986	0.999	0.999
16	0.124	0.205	0.356	0.563	0.906	0.994	0.999	0.999
18	0.127	0.220	0.393	0.618	0.939	0.998	0.999	0.999
20	0.130	0.236	0.429	0.668	0.962	0.999	0.999	0.999
25	0.139	0.275	0.516	0.774	0.988	0.999	0.999	1.000
30	0.147	0.315	0.596	0.852	0.997	0.999	1.000	1.000
35	0.156	0.356	0.668	0.906	0.999	0.999	1.000	1.000
40	0.165	0.396	0.732	0.942	0.999	0.999	1.000	1.000
50	0.183	0.475	0.830	0.980	0.999	1.000	1.000	1.000

Source: fpower.sas macro retrieved from http://www.math.yorku.ca/SCS/Online/power/ on 1 March 1 2005.

APPENDIX C

Sampling Strategies

C-1. Introduction. As addressed in USACE's Technical Project Planning—Phase I, project technical staff must consider which sampling strategy is appropriate for the current project phase (EM 200-1-2). It is not necessary to apply the same strategy throughout all phases of a project's life cycle. Frequently, early screening sampling may employ a simple strategy, and subsequent phases may require more complicated strategies, using data results from previous phases. Whenever possible, it is best to use available site knowledge in developing a sampling strategy.

C-1.1. Although there are many sampling approaches, this Appendix presents a discussion of the most commonly employed strategies, which are:

a. No sampling.

b. Judgmental sampling.

c. Random sampling.

(1) Simple random sampling.

(2) Stratified random sampling.

(3) Systematic and grid sampling.

d. Ranked set sampling.

e. Composite sampling.

f. Adaptive sampling.

C-1.2. The first two strategies are qualitative; the remaining strategies are probabilistic. In the latter, statistics may be used to estimate sample characteristics such as mean, standard deviation, and uncertainties. Whether performing on-site, field, or off-site laboratory analysis, the sampling design requires equal consideration. For further insights into environmental sampling, see Gilbert (1987) and EPA/600/R-96/084.

C-2. No Sampling. It may be possible to establish the absence of human health or environmental risk without any sampling. There are three criteria necessary to create a quantifiable risk: i) a chemical release to the environment; ii) a pathway of exposure; and iii) an exposed population. If any of these conditions are not satisfied, a risk does not exist and sampling is not required.

C-2.1. Historical quantitative and qualitative information available during the early stages of a project's life cycle may be adequate for site closure without sampling. Qualitative data are typically not as expensive to collect as quantitative data and may be more informative than quantitative data for answering questions about hazardous, toxic, and radioactive waste sites.

C-2.2. Historical qualitative and quantitative data hold an array of site information useful in reaching a conclusion. The reliability and applicability of historical data and qualitative information (such as interviews with site personnel and photographs) should be evaluated. For example, have historical chemical data been gathered using comparable methods? Is the set of material safety data sheets complete and current? Does toxicity data derived from studies demonstrate adequate quality control? Are engineering drawings pre- construction or “as-builts”? Statistical techniques are often critical to assessing the usability of quantitative historical data, particularly when incorporating historical data into more re- cent data sets. Simple descriptive statistics (such as the mean, standard deviation, and range) and statistical plots (such as box-and-whisker plots) are useful for qualitative comparisons of different data sets (Appendices D and J). Quantitative statistical comparisons are also frequently appropriate. For example, it may be desirable to compare the mean or variance of a prior data set to a recent data set (Appendix M). When quantitative statistical comparisons are made, the data should also be evaluated to verify that they satisfy the underlying assumptions of the statistical tests (for example, random sampling and adequate numbers of samples).

C-3. Judgmental Sampling. Perhaps the most common sampling strategy is judgmental sampling (also known as targeted or biased sampling). As the name implies, this sampling strategy relies upon the investigator’s knowledge and experience. Judgmental sampling is the selection of samples without a statistical design, that is, without any randomization. It can be useful when good documentary data are available and when it is done by an experienced professional with technical expertise. Judgmental sampling is frequently used to target high-contaminant concentrations or worst-case site conditions, such as the collection of samples in visibly stained soils. The underlying rationale for this approach is that, if contamination were not detected (or detected at acceptable levels) in the areas of the site that would have been most impacted by site-related waste handling activities, then acceptable levels of contamination could be assumed in the remaining portions of the study site. However, if unacceptable levels of contamination were detected, the results would be inappropriate for evaluating site-wide average concentrations. An example of judgmental sampling is presented below to illustrate a common improper use of the sampling technique.

C-4. Case Study 1—Judgmental Sampling, Ordnance Demolition Area.

C-4.1. The project team used judgmental sampling to obtain a worst-case estimate of explosive residues in surface soils associated with an ordnance demolition area. They did this by sampling where activities historically occurred, specifically targeting stained soils, pits, and debris-laden areas. The team collected background samples and compared group means and variances. They found a statistically significant increase in on-site concentrations relative to the background samples for several explosive residues, concluded that the entire site was contaminated with explosives, and scheduled the area for further investigation and remediation.

C-4.2. In this case, it was incorrect for the project team to compare judgmental non-randomized data sets in a statistically quantitative manner. This problem is common in using historical data. One of the primary assumptions in conducting any statistical analysis is that data were obtained in a random fashion. The fact that the on-site samples were biased toward areas of known or suspected high concentration increased the probability that the on-site average concentration would exceed background, potentially leading to biased conclusions. Either the initial round of sampling should have been performed randomly or new samples should be randomly collected and submitted for analysis prior to concluding the presence of site-wide contamination. Alternatively, it might be possible to stratify the site in such a manner that the judgmental samples are representative of only select portions of the entire study area. See Section II of Chapter 3 for further discussion of comparing on-site to background concentrations.

C-5. Random Sampling. The term random sampling encompasses a set of unbiased techniques to choose locations from which to sample at a site. Random sampling has the advantage that its lack of bias allows for robust statistical calculations. However, random sampling is not the same as arbitrary sampling; it does not mean "sample in any manner." The sampling design must be such that every portion of the population possesses an equal opportunity of being selected in the sample. Therefore, when implementing a random sampling design, planners must define and consider the entire population. Both the spatial and temporal boundaries of the environmental population must be well-defined, as instructed in EPA/600/R-96/055, QA/G-4. Samples may need to be collected randomly, not just horizontally across a study area, but vertically as well. Likewise, a continuing waste stream would be sampled randomly in time. Three forms of random sampling are discussed in this paragraph: simple random sampling, stratified random sampling, and systematic random sampling. EPA Quality Assurance QA/G5-S, Guidance for Choosing a Sampling Design for Environmental Data Collection, describes the three random sampling methods in detail.

C-5.1. Simple RandomSampling. In simple random sampling, sample locations are selected using random numbers. Every possible set of locations has an equal chance of being selected. For example, a simple random sample from a group of liquid waste drums may be taken by numbering all the drums and randomly selecting numbers from that list. Simple random sampling does not presuppose any information regarding the spatial distribution of the likely contamination at the site, other than assuming that no spatial correlation exists. Samples are collected at random from the study area without consideration for factors such as suspected disposal activities, debris locations, spills, or other spatial control on contamination.

C-5.1.1. The major advantages of simple random sampling are that i) it provides statistically unbiased estimates of the mean, proportions, and variability; ii) it is easy to understand and use; and iii) sample size calculations and data analysis are simple to do.

C-5.1.2. The disadvantages of simple random sampling are as follows.

C-5.1.2.1. The environmental population must be relatively homogeneous for simple random sampling to be effective. In particular, major spatial or temporal trends should not exist. Simple random sampling would be inappropriate if localized areas of high contamination or hot-spots exist. Because every portion of the site has an equal opportunity of being selected, if hot-spots constitute a small portion of the total study area, it is likely that random sampling will fail to detect them. Under these circumstances, random sampling will give undue weight to the less contaminated portions of the site.

C-5.1.2.2. It is possible that, by random chance alone, the sample points will be clustered within a small portion of the study area and will not reliably characterize (e.g., owing to heterogeneity) the entire study area.

C-5.1.2.3. Random sampling is often less efficient and, as a result, more expensive than other sampling designs because it requires more samples to obtain the same result. It is most viable when the target population or study area is small. The analytical costs may be offset by the streamlined sampling design, which requires less research than judgmental sampling.

C-5.2. Stratified Random Sampling.

C-5.2.1. In stratified sampling, the target population is separated into non-overlapping sub-populations, or strata, that are expected to be relatively homogeneous. Strata may be chosen on the basis of spatial or temporal proximity of the units or on the basis of existing information or professional judgment about the site or process. For instance, if an exposed population is likely to contact only surface soil rather than all soil, then the site could be divided into a surface soil stratum and subsurface soil stratum. Once the strata are defined, each stratum is randomly sampled. This approach allows the project team to focus on areas of greatest concern while retaining the benefits of a random sampling plan. Some examples of stratification at a hazardous waste site include different soil types, depth within an aquifer or surface water body, or separate waste ponds used at different times in site history.

C-5.2.2. Stratified random sampling can be a very effective approach to site characterization. If there is less variation within each subpopulation than in the target population as a whole, stratified random sampling can be more efficient than simple random sampling. Other advantages of this design are that it has potential for achieving greater precision in estimates of the mean and variance, and that it allows computation of reliable estimates for population subgroups of special interest. In fact, a well-constructed stratified sampling plan is the best alternative in most instances where judgmental sampling plans are now employed.

C-5.3. Systematic RandomSampling. In systematic sampling, samples are taken at regular intervals in time or space, i.e., along some sort of grid. An initial location or time is selected at random, and subsequent samples are collected at regular spatial or temporal intervals. The sampling scheme retains its random characteristic as long as the initial sampling location or time is randomly, not arbitrarily, selected.

C-5.3.1. Systematic sampling methods are used to search for hot-spots and to infer means, percentiles, or other parameters. They are also useful for estimating spatial patterns or trends over time. These designs provide practical and easy methods for designating sample locations and ensure uniform coverage of a site, unit, or process. One significant benefit of a systematic design is that it generally ensures that some samples from each possible sub- group within a population will be selected.

C-5.3.2. There are two approaches to grid sampling. One may select a particular grid pattern and sample at every node within the grid. Although it is common for sampling plans to specify a square grid pattern, there are a variety of patterns that can be used, often to some advantage in terms of cost or efficacy. Grid blocks may be squares, rectangles, triangles, parallelograms, pentagons, hexagons, or other polygons, depending upon the application. Alternatively, one may randomly pick a starting point in a grid and then collect samples in some logical pattern (for example, move south two blocks and east three blocks). When the edge of the grid is encountered, the pattern starts again on the opposite side of the grid.

C-5.3.4. One can immediately see that such an approach could be very expensive. This type of sampling is often reserved for situations where the analytical cost is low, or where the area to be covered is quite large, as in the estimation of lead analysis over a firing range using a portable x-ray fluorescence (XRF) spectrometer. An important consideration is the size of the individual blocks within the grid or the distance between grid lines.

C-5.4. Hot-Spot Sampling. Searching for a hot-spot is a special case where grid spacing may be estimated using information about the suspected hot-spot size and shape. Hot-spots may be located on two-dimensional surfaces or in three-dimensional volumes. For volumes, a three-dimensional grid is generated via the extension of a pair of two-dimensional grids.

C-5.4.1. This method relates the likelihood of successfully locating hot-spots based on their assumed size, shape, and orientation. The acceptable probability of not finding a hot-spot (β) must be specified at the outset. This value must be decided upon by the project team depending on the degree of risk associated with not identifying the hot-spot. Gilbert (1987) provides graphs (called nomographs) that correlate the shape of the hot-spot with the acceptable probability of not finding the spot and the length of the hot-spot divided by the required grid spacing. Table C-1 provides a summary of the nomographs for square and triangular grids. Users will need to interpolate, reference the original citation, or use a conservative set of values in applying this table to individual studies.

C-5.4.2. As mentioned above, to determine the grid spacing (*G*) for a hot-spot, assumptions must be made about its size and shape (Figure C-1). The shape is represented by the factor (*S*), defined as the width (*W*) of the elliptical target spot divided by the expected length (*L*). If the expected shape is a circle, *S* is equal to 1. If *S* is an ellipse, *S* is less than 1, but greater than 0. If *S* is unknown, planners may choose to assume that the hot-spot is a narrow elliptical shape, i.e., *S* is 0.5 or less. This assumption is conservative. Accommodating a narrower target shape results in denser grid spacing.

Table C-1.
Hot-Spot Grid Spacing

For Square Sampling Grids—Values Listed Are *L*/*G*

	S									
β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0						1.00	0.80	0.77	0.74	0.70
0.1				1.00	0.83	0.74	0.68	0.62	0.58	0.55
0.2				0.87	0.77	0.68	0.62	0.58	0.53	0.51
0.3			0.93	0.78	0.69	0.62	0.57	0.53	0.49	0.47
0.4			0.85	0.72	0.64	0.58	0.53	0.49	0.47	0.44
0.5		0.94	0.77	0.65	0.57	0.51	0.48	0.44	0.42	0.40
0.6		0.83	0.68	0.58	0.51	0.47	0.43	0.41	0.39	0.37
0.7	1.00	0.71	0.58	0.50	0.44	0.41	0.38	0.35	0.33	0.31
0.8	0.78	0.56	0.44	0.49	0.35	0.32	0.30	0.28	0.27	0.26
0.9	0.57	0.39	0.32	0.29	0.27	0.25	0.23	0.21	0.20	0.19
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

For Triangular Sampling Grids—Values Listed Are *L*/*G*

	S									
β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0					0.94	0.81	0.74	0.66	0.60	0.57
0.1				0.90	0.78	0.69	0.62	0.57	0.52	0.50
0.2			0.95	0.80	0.70	0.62	0.57	0.52	0.49	0.47
0.3			0.87	0.73	0.63	0.57	0.52	0.48	0.46	0.43
0.4		1.00	0.79	0.67	0.58	0.53	0.48	0.45	0.42	0.40
0.5		0.86	0.69	0.59	0.52	0.48	0.43	0.41	0.39	0.37
0.6		0.75	0.61	0.52	0.47	0.42	0.39	0.37	0.35	0.32
0.7	0.94	0.84	0.52	0.44	0.40	0.37	0.33	0.31	0.30	0.28
0.8	0.75	0.52	0.41	0.37	0.32	0.30	0.28	0.27	0.24	0.22
0.9	0.51	0.36	0.30	0.25	0.22	0.20	1.90	1.80	1.70	1.70
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

β = probability of missing the hot-spot

W = width of elliptical hot-spot
L = length of the semi-major axis (radius of a circle)
G = grid spacing

S =*W*/*L* (ratio of width to length of hot-spot)
S = 1 is a circle
S = 0.1 is a narrow ellipse
L/*G* = a dimensionless value

C-5.4.3. Based on an estimate of the length of the target hot-spot, we may define the value (L), which is one-half of the long axis of the ellipse. In the case of a circular hot-spot (S = 1), this is equivalent to the radius of the circle. Finally, the nomographs presented as Table C-1 may be used to determine the appropriate grid spacing (expressed in terms of L/G), based on the values of S and β.

C-5.4.4. The effectiveness of the hot-spot sampling method depends on the accuracy of existing site-specific information. Without prior knowledge, it is difficult for planners to es- timate the shape and dimensions of the anticipated hot-spot. In practice, this information is rarely known with confidence, and hot-spot spatial dimensions are often determined on the basis of economic considerations rather on the basis of pre-existing information on site conditions. The required number of samples depends greatly upon the assumed dimensions of the hot-spot. Planners should do a number of calculations, varying the shape and estimated size of the hot-spot. If the resulting grids are similar and differences in sample design relatively minor, then planners may feel more confident about the methodology applied to the site.

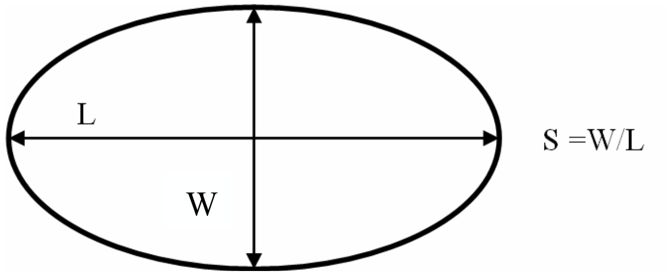


Figure C-1. Elliptical Hotspot.

C-5.4.5. The hot-spot mathematical procedure may also be applied in reverse; if grid spacing and presumed hot-spot size and shape are known, the probability of having missed a hot-spot (of some specified size) may be determined. Thus, site investigation managers may be able to convey to regulators the level of certainty that no problems were missed, within reasonable expectations. By applying the nomographs and solving for different variables, a researcher can answer such questions as the size of a hot-spot likely to be found by a given grid spacing, and the probability of not finding a hot-spot based on a given grid spacing. The following case study compares sampling strategies for a site with a hot-spot.

C-6. Case Study2—Comparing Random Sampling Strategies at a Site with a Hot-Spot. Table C-2 illustrates examples of the three random sampling approaches at a generic site and the differences in descriptive statistics that might influence a manager's decisions related to the site. The three different sampling plans are applied to the same data set: Plan A is simple random sampling, Plan B is stratified random sampling, and Plan C is systematic and grid sampling. The site is represented by a 9-by-9 grid with the 3 right-most grid columns divided by a heavy solid line indicating a hot-spot, and the lower left 12 cells a secondary hot-spot (applicable to Plans B and C only). For Plans B and C the largest group of cells is Group 1; the lower left corner is Group 2; and the right three columns make up Group 3. The number in each cell represents a generic analytical result, had a sample been collected from every cell. A collected sample is represented by a shaded cell. The systematic sampling (Plan C) was determined by using a set pattern beginning at a randomly selected first location. (This is not obvious from the pattern of shaded cells.)

C-6.1. For this example, assume that decisions will be based on a 2-stage comparison criterion: values less than 5 units require no action; values greater than 5 units but less than 50 units require further remedial investigation but no immediate action; and values greater than 50 units indicate an immediately dangerous condition requiring an emergency removal action.

C-6.2. The three sampling plans are judged against a hypothetical sampling of every cell across the site. In this case, the following are determined:

C-6.2.1. Total number of samples, $N = 81$.

C-6.2.2. Summation of all results, $S = 708.1$.

C-6.2.3. Total population average, $\mu = 21.09$.

C-6.3. For Plans B and C, the following are determined for the entire populations of each group:

Group 1:	$n_1 = 42$	$S_1 = 23.4$	$\mu_1 = 0.56$
Group 2:	$n_2 = 12$	$S_2 = 47.7$	$\mu_2 = 3.98$
Group 3:	$n_3 = 27$	$S_3 = 1,637$	$\mu_3 = 60.63$

C-6.4. Note that population mean may be viewed as a weighted mean calculated from each group population mean:

$$\mu = \sum_i (n_i / N)\, \mu_i = \sum_i w_i \mu_i$$

$$\mu = \sum_i w_i \mu_i$$

$w_1 = 42/81$, $w_2 = 12/81$, $w_3 = 27/81$

C-6.5. For Plans B and C, a total of nine samples are randomly selected. (For example, for the nine samples collected for Plan B, two are from Group 1, four are from Group 2, and three are from Group 3.) The mean of the population mean (i.e., entire set of 81 samples) is estimated by calculating the sample mean of each group and weighting them:

$$\bar{x} = \sum (n_i / N)\,\bar{x}_i = \sum_i w_i \bar{x}_i$$

C-6.6. To assess each sampling plan, the mean concentrations determined from the limited sampling to those for the entire site data set are compared. Simple random sampling (Plan A) provides the best estimate of the overall population average. However, it is fairly limited in identifying the best course of action for the underlying strata in that it suggests that the entire population is subject to additional investigation or action. Another shortcoming is that none of the random sampling designs identified the "secondary hot-spots" in Group 2; that is, none of the samples selected in Group 2 (the shaded cells) exceed 5. Stratified sam- pling (Plan B) resulted in better data for decision-making because data were obtained for all three groups, although some of the group mean estimates are rather poor. In the systematic plan (Plan C), each stratum is represented in the statistics at a frequency roughly equal to its portion of the whole. (The ratio of the total number of cells for Groups 1, 2, and 3 is approx- imately 5:1:3, the ratio of the number of samples collected for each group.) Had the presence of underlying strata been unknown, the systematic plan would have given the best indication of potential problems at the site.

C-7. Systematic SamplingOverTime. Systematic sampling can also be applied when the parameter of interest is expected to vary over time. This one-dimensional scheme is sometimes called periodic sampling and is quite simple. Divide the span of time under examination into an arbitrary number of "blocks" (e.g., 20 intervals) and, having calculated an appropriate number of samples for the application, simply divide the number of samples required into the number of blocks available. This gives the time between samples. The starting time is chosen randomly. (Note that the same strategy may be used to establish the distance between grid lines, where the intervals would be measured in units of distance rather than time.) In general, the greater the variability in the parameter being measured is, the greater the number of samples required for the required degree of confidence.

C-8. Ranked Set Sampling. As stated in EPA QA/G5-S: "Ranked set sampling is an innovative design that can be highly useful and cost-efficient in obtaining better estimates of mean concentration levels in environmental media." The technique typically entails the use of two analytical methods, a "definitive" method (e.g., a fixed laboratory method) and a "screening" method (e.g., a field method). Usually, the cost of the screening method is significantly less than that of the definitive method, while the analytical quality of the definitive method exceeds that of the screening method. Ranked set sampling is a two-phase sampling design. It first identifies sets of field locations and uses inexpensive measurements to rank locations within each set; next, it selects one location from each set for analysis by the definitive method. Only a brief overview of this sampling technique is presented in this Appendix. The reader is referred to the EPA QA/G5-S guidance document for a more detailed discus- sion and illustration of rank set sampling.

C-8.1. For a “balanced design,” *m* sets of *m* samples (at total of m^2 samples) are initially analyzed using professional judgment or some screening method. The field samples in each set are then independently ranked (e.g., from highest to lowest). The first ranking sample (the highest sample) is selected from the first set, the second highest ranking sample is selected from the second set, and so forth, until *m* samples are selected for analyses using the definitive (i.e., more accurate and expensive) analytical method. The process is repeated *r* times, giving a total of $m^2 r$ field analyses and *mr* definitive analyses.

C-8.2. One of the best reasons for applying ranked set sampling is its ability to provide samples from across the distribution of values at the site. This, in turn, creates a better estimate of the population mean and improves the performance of various other statistical tests, especially those that entail distributional assumptions. A wide variety of field screening tools can be used to supplement the professional judgment of the samplers and, in certain circumstances, can even be used later as definitive data, assuming good correlation with fixed laboratory results is achieved. Paragraph C-9 illustrates a practical application of ranked set sampling.

C-8.3. Relative to simple random sampling, this design results in a more representative sample, and therefore leads to more precise estimates of the population parameters. A large number of screening analyses increases site coverage, and the ranking information from the screening analyses reduces the required number of definitive analyses relative to the number that would be required from a random sampling design. Therefore, the ranked set sampling approach has the added benefit of typically being less expensive than a simple random sampling approach. Because preliminary data are used to ensure representative samples are collected, the variability among the samples is better controlled and the number of samples required to make a probabilistic decision with the same degree of confidence is reduced.

C-8.4. However, there are several limitations to ranked set sampling. The screening and definitive methods must be strongly correlated with one another. In addition, the cost of the definitive analyses compared to the cost of the ranking procedure used for the field methods must be relatively large for the approach to be cost-effective. One should consider whether two phases of sampling is cost-effective relative to a more standard sampling method and whether it is technically feasible given project resource constraints. Finally, the statistical computations to be performed on the resulting data set are more complex relative to those used for a simple random sampling design.

Table C-2. Comparison of Random Sampling Method Results

Plan A: Simple Random Sampling

0.26	0.24	0.74	0.95	**0.25**	0.34	**94.18**	20.16	61.90
0.97	0.54	0.13	0.18	**0.17**	0.48	5.40	13.39	**19.79**
0.97	0.30	0.72	0.09	0.48	0.79	55.28	55.10	94.98
0.82	0.03	0.95	0.72	**0.22**	0.81	29.31	1.26	72.37
0.52	0.66	**0.48**	0.83	0.92	0.43	78.73	84.02	**77.05**
2.82	1.45	1.24	0.52	0.69	0.47	89.00	98.76	83.54
3.14	8.24	8.48	0.55	0.11	0.85	76.71	96.91	84.19
7.18	1.68	0.96	**0.74**	0.47	0.86	42.95	16.94	72.67
5.84	3.73	2.98	0.65	0.99	0.51	96.66	52.85	62.86

Plan A	S_i	$\bar{x}_i$	$\bar{x} = \sum w_i \bar{x}_i$
All nine samples	193.84	21.54	N/A

Plan B: Stratified Random Sampling

0.26	0.24	0.74	0.95	**0.25**	0.34	94.18	20.16	61.90
0.97	0.54	0.13	0.18	0.17	0.48	5.40	**13.39**	19.79
0.97	**0.30**	0.72	0.09	0.48	0.79	55.28	55.10	**94.98**
0.82	0.03	0.95	0.72	0.22	0.81	29.31	1.26	72.37
0.52	0.66	0.48	0.83	0.92	0.43	78.73	84.02	77.05
2.82	**1.45**	1.24	0.52	**0.69**	0.47	89.00	98.76	83.54
3.14	8.24	8.48	0.55	0.11	0.85	76.71	96.91	84.19
7.18	**1.68**	0.96	0.74	0.47	0.86	42.95	16.94	72.67
5.84	3.73	2.98	0.65	0.99	0.51	**96.66**	52.85	62.86

Plan B	S_i	$\bar{x}_i$	$\bar{x} = \sum w_i \bar{x}_i$
All nine samples	210.22	23.36	N/A
Group 1	2.06	0.51	
Group 2	3.12	1.56	
Group 3	205.04	68.35	
			23.28

Plan C: Systematic and Grid Sampling

0.26	0.24	0.74	0.95	0.25	0.34	94.18	20.16	61.90
0.97	0.54	0.13	0.18	0.17	0.48	5.40	13.39	19.79
0.97	**0.30**	0.72	**0.09**	**0.48**	0.79	55.28	55.10	94.98
0.82	0.03	0.95	0.72	0.22	0.81	29.31	1.26	72.37
0.52	0.66	0.48	0.83	**0.92**	0.43	78.73	84.02	77.05
3.14	8.24	8.48	0.55	0.11	0.85	76.71	**96.91**	84.19
7.18	1.68	0.96	0.74	0.47	0.86	42.95	16.94	72.67
5.84	**3.73**	2.98	0.65	0.99	0.51	96.66	**52.85**	62.86

Plan C	S_i	$\bar{x}_i$	$\bar{x} = \sum w_i \bar{x}_i$
All nine samples	244.75	27.19	N/A
Group 1	2.26	0.45	
Group 2	3.73	3.73	
Group 3	238.76	79.59	
			27.31

SUMMARY

Grouping	Population Mean#	Simple $\bar{x}_A$	Stratified $\bar{x}_B$	Systematic $\bar{x}_C$ #
Group 1	0.56	—	0.52	0.45
Group 2	3.98	—	1.56	3.73
Group 3	60.63	—	68.35	79.59
Entire Grid	21.09	21.54	23.28	27.31

Notes:
Shading indicates a sampled grid location

C-9. Case Study3—RankedSet Sampling. The project team used field screening test kits on a grid established over a wide area to characterize an ordnance demolition area. Using the information from the field screening, the team was able to stratify the site into three areas: i) a region requiring no remediation; ii) an area clearly requiring remediation and for which samples at depth were required to provide volume estimates; and iii) an area requiring additional study with definitive methods to establish the need for remediation or no further action. Definitive samples were then collected to distinguish the various explosives and their daughter products that the test kit could not resolve. These results were then used to better estimate the average concentration of individual explosives within the various strata, and to serve as confirmation samples for the test kits. The definitive samples helped correlate low-, mid-, and high-range concentrations in each area. Thus, the screening data were used to select locations for definitive samples to ensure more representative mean concentrations within each area.

C-10. Composite Sampling. Composite sampling is the physical averaging of environmental samples in a manner that yields an accurate and representative estimate of environmental conditions, usually at a reduced cost. It involves physically combining and homogenizing two or more environmental samples (referred to as "grab" samples, and called "subsamples" in this context) to form a new sample referred to as a composite sample. Compositing is used when the mean is primarily of interest (i.e., because the process is a physical averaging) and information on the spatial or temporal variability of contamination is not needed (i.e., because this information is lost unless the subsamples can be reanalyzed). Tables C-3 and C-4 suggest circumstances under which compositing can be useful. Various sampling designs may be used to select subsamples to be mixed together into composites.

Table C-3.
Objectives of Composite Sampling—Fundamental Cases

1. Objectives that rely on composite sampling	a. Estimating a population (or stratum) mean for a continuous variable (e.g., analyte concentration)*****
	b. Estimating proportion of population exhibiting some trait
2. Objectives that rely on composite sampling and retesting protocols	a. Classifying sampling units as having or not having some trait such as being in a hot-spot or from a contaminated cell
	b. Identifying the sampling unit with highest value of some continuous measure (e.g., concentration), or identifying sampling units in the upper percentiles
* In general, information on variability and spatial or temporal patterns is lost when compositing is used for this objective; however, in some cases, some information on patterns can be acquired.	

Table C-4.
Criteria for Judging Benefits of Composite Sampling

Criterion or Objective	Composite sampling is likely to be beneficial if…
1. Analytical costs	Analytical costs are high relative to sample acquisition/ handling costs.
2. Analytical variability	Analytical variability is small relative to variability of the target population.
3. Analytical sensitivity	Concentrations of relevance are much larger than detection and quantitation limits.
4. Representativeness	Compositing does not affect sample integrity (expect no chemical reactions/interferences or analyte losses from volatility) or result in safety hazards. Individual samples can be adequately homogenized.
5. Objective is to estimate population mean (See 1a in Table 2-3)	Information on individual samples is not important. Information on associations is not important. Criteria 1, 2, and 4 are met.
6. Objective is to estimate proportion of population with a trait (See 1b in Table 2-3)	Composite has trait if individual sample does. Likelihood of misclassification is small. Trait is rare. Criteria 1, 2, 3, and 4 are met.
7. Objective is to classify samples as having/not having a trait (See 2a in Table 2-3)	Composite has trait if individual samples do. Likelihood of misclassification is small. Retesting of aliquots (grab samples) for each composite sample is possible. Trait is rare. Criteria 1, 2, 3, and 4 are met.
8. Objective is to identify the sample(s) with the highest value (See 2b in Table 2-3)	Measurement error is negligible. Retesting of aliquots from individual samples is possible. Criteria 1, 2, 3, and 4 are met.

C-11. Compositing Fluids. A typical application of compositing fluids is in creating a repre sentative sample when one or another condition, tied to contaminant mass or concentration, varies over space or time. National Pollutant Discharge Elimination System (NPDES) monitoring provides a classic case in point.

C-11.1. The fundamental objective for this type of compositing is to develop a single sample that accurately represents the whole area or time under consideration. The alternative entails greatly increased sampling and analysis costs and agreement on an acceptable mathematical approach to combining the individual sample results. Table C-5 examines a variety of compositing approaches linked to particular circumstances. Paragraph C-12 illustrates an example of flow-proportioned compositing.

C-11.2. Another classic use of compositing fluids is in sampling stack emissions. When a fluid (or gas in the case of stack emissions) flows through a pipe, the fluid does not move at a uniform speed across the diameter of the pipe. Friction with the interior surface of the pipe causes fluids near the casing to move more slowly than at the center. Thus, when measuring mass per unit volume per unit of time, isokinetic sampling is applied. In this case, subsamples are collected across the diameter of the pipe for identical time intervals, along with a measure of the flow rate at the individual locations. Using this information, the engineer can balance concentration against the flow rate to yield an accurate estimate of the average mass discharged from the stack (or pipe) over time.

Table C-5.
Compositing Methods

Method No.	Sampling Mode	Compositing Principle	Comments	Disadvantages
1.	Continuous	Constant sample pumping rate	Practicable but not widely used	Yields large sample volume; may lack representativeness for highly variable flows
2.	Continuous	Sample pumping rate proportional to stream flow	Not widely used	Yields large sample volume but requires accurate flow measurement equipment
3.	Periodic	Constant sample volume, constant time interval between samples	Widely used in automatic samplers and widely used as manual method	Not most representative method for highly variable flow or concentration conditions
4.	Periodic	Constant sample volume, time interval between samples proportional to stream flow	Widely used in au tomatic sampling but rarely used in manual sampling	Manual compositing from flow chart
5.	Periodic	Constant time interval between samples; sample volume proportional to total stream flow since last sample	Not widely used in automatic samplers but may be done manually	Manual compositing from flow chart
6.	Periodic	Constant time interval between samples; sample volume pro- portional to stream flow at time of sampling	Used in automatic samplers and widely used as manual method	Manual compositing from flow chart

After: EPA 600/4-82-029

C-12. Case Study 4—Flow-Proportioned Compositing. At a manufacturing facility in Ohio, an existing NPDES permit called for the facility to collect a single, three-part, equal-weight composite sample monthly. The facility operated three shifts. Production on all three shifts was essentially the same, although the bulk of maintenance activities took place on the second shift. Three grab samples, one from each shift, were composited at the laboratory prior to analysis.

C-12.1. A change in business climate led to a reduction in demand such that the midnight to 8 a.m. shift was canceled and the 4 p.m. to midnight shift was reduced by roughly two-thirds. The facility manager asked that the overall effect the change in shifts would have on discharge rates be assessed in preparation for permit renewal negotiations. For this case study, only the nitrate data are considered. The following analysis was performed:

Original flow–shift 1	200,000 gal/day[*]	New flow	200,000 gal/day
Original flow–shift 2	200,000 gal/day	New flow	70,000 gal/day
Original flow–shift 3	200,000 gal/day	New flow	5,000 gal/day

C-12.2. Historical composite results for the previous year were as follows:

Jan	0.48	Average	0.38 mg/L[†]
Feb	0.12	Variance	0.20 mg/L
Mar	0.26		
Apr	0.34	Current Permit Limit	2.5 lb/day[‡]
May	0.48	EPA Proposed New Limit	1.0 lb/day
Jun	0.31		
Jul	0.47		
Aug	0.46		
Sep	0.13	Assuming average concentration does not change	
Oct	0.40		
Nov	0.16	Under Equal Volume sampling, lb/day = 1.9	
Dec	0.20	Under Flow Proportioned sampling, lb/day = 0.87	

C-12.3. Thus, the new permit limit will be acceptable if the permit also incorporates a change in the compositing method.

[*] gal/day = gallons per day
[†] mg/L = milligrams per liter
[‡] lb/day = pounds per day

C-13. Compositing Solids. Generally speaking, solids and, in particular, soils are composited to estimate the concentration of a contaminant over large areas, or when the granular or globular nature of the contaminant of concern (e.g., explosives, PCB oils) can provide false estimates of concentration from individual measurements because of excessive heterogeneity in the individual samples. Other applications are also possible. Compositing can also be used to assess the proportion of samples that meet a specific condition and, with retesting of a small subset of original locations, can also be used to locate rare events (like hot-spots) where too many individual samples would be required. For example, at a site with very few historical data, 12 composite samples of 4 subsamples each may be analyzed for a long list of possible contaminants. If only one sample contains only a few contaminants of concern, then further investigation is limited to those contaminants and in only four small areas. Exhaustive testing of the 48 original discrete samples was not necessary, and further study of most of the site is precluded. As extensive mixing of the subsamples is required to form a representative composite, composite sampling is not generally applied to samples when volatile organic compounds (VOCs) are of particular interest.

C-14. Adaptive Sampling. Adaptive sampling designs are typically used to characterize the extent of contamination using multiple sampling events; they rely upon cost-effective field methodologies with rapid turn-around time. The results of an initial sampling event are used to modify the selection of future sampling locations for the study area. Adaptive cluster sampling is useful when the characteristic of interest is sparely distributed through the site. Adaptive cluster sampling could be used for a study area that contains mostly low-level or negligible contamination but also isolated pockets of high-level contamination (i.e., hot-spots). This is illustrated in Figure C-2. As stated previously, under these circumstances, a random sampling design would not be the optimum approach (as the hot-spots could remain undetected).

C-14.1. Three major elements characterize adaptive cluster sampling. First, a set of sampling locations is initially determined. Though there may be insufficient data to support firm conclusions overall, information may exist that suggests particular areas of the site are clean or contaminated. The result is an initial conceptual model for the site. For example, a grid is placed over the geographical area of interest, where each cell of the grid represents a potential sampling unit (location). A subset of all the potential sampling units is selected for sampling. Figure C-2 illustrates the use of random sampling for the selection of the initial sampling event. Second, a decision rule for each sampling unit must be established. If the contaminant of interest exceeds the decision limit, additional sampling is required “near” the sampling unit (i.e., adjacent sampling units are sampled). Third, the “neighborhood” of each sampling point (i.e., the area required for additional sampling) must be defined. Several additional stages of sampling are designated on Figure C-2. The symbol “X” denotes the neighboring sampling units that were sampled. (Note: In the example illustrated in Figure C-2, one area of contamination was missed.) The decision rule and additional sampling are repeatedly applied until contamination is not detected above the decision limit for each sampling unit. This results in a “mapping” of contaminants as illustrated in the final stage in Figure C-2, where the extent of “hot-spots” is delineated using a large number of sample units. The shaded areas in Figure C-2 represent “hot-spots” (i.e., area in which contamination exceeds the decision limit).

C-14.2. Adaptive sampling and analysis plans (SAPs) provide a cost-effective alternative to traditional sampling designs. Adaptive SAPs are based on field analytical methods allowing for rapid sample turnaround and field-based decision support to guide the sampling program. One objective of adaptive SAPs is to support removal actions.

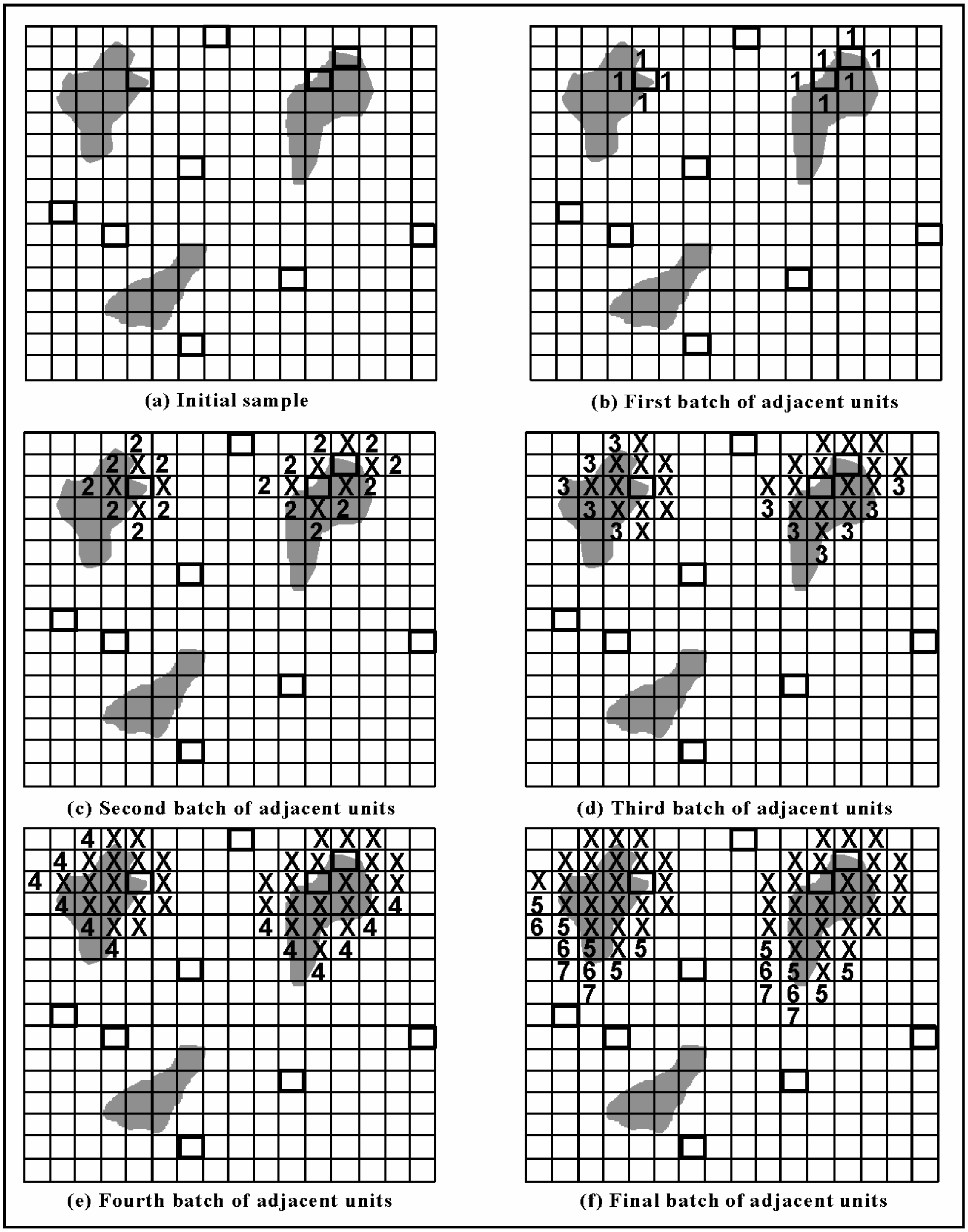
C-14.3. Traditional approaches to designing and executing a removal action have relied on "digging to the design line" and then taking confirmation samples. The static work plans that have accompanied these efforts have specified the number and location of samples. Often, however, the design lines have been at best rough approximations of the real extent of contamination, resulting in either extensive under- or over-removal of soils. In both cases, the economic impacts have been significant. An important factor in establishing the design line is the site cleanup levels. Cleanups should be implemented so that concentrations left at the site meet the cleanup goal to a predetermined level of certainty, with the level of certainty agreed upon by the design team and regulators.

C-14.4. Adaptive SAPs rely on field analytical methods to generate sample results quickly enough to have impact on the course of the sampling program. They are based on dynamic work plans that specify the logic of how sampling numbers, locations, and analyses will be determined as the program proceeds. They also rely on rapid, field-level decision-making. Adaptive SAPs require: i) field analytical methods that are appropriate for the types of contaminants expected at a site; and ii) a means for supporting decision-making in the field that is appropriate for the goals of the program.

C-14.5. Rapid field decision-making requires qualitative and quantitative decision support. Qualitative decision support means having technical staff equipped with an accurate understanding of the sampling progress. Large adaptive SAPs can produce hundreds of samples per day. Managing, integrating, and displaying the sample information pose a serious logistical challenge that can interfere with program process if not adequately addressed. A typical adaptive SAP includes some type of field- or web-based database system along with a Geographic Information System for data display to help with logistics and visualization.

C-14.6. Quantitative decision support for adaptive SAPs that delineate removal areas requires the ability to estimate contaminant extent based on sampling results, determine the uncertainty associated with those results, predict expected values from previous sampling, and identify new removal locations based on that information.

C-14.7. The adaptive sampling scheme presented in Figure C-2 may be applied to contamination removal actions as well. In such an application, each sample is used to determine whether soil removal (i.e., excavation) is necessary, and the areal (and volumetric) extent of soil needing removal can be established via such sampling techniques.



(a) Initial sample

(b) First batch of adjacent units

(c) Second batch of adjacent units

(d) Third batch of adjacent units

(e) Fourth batch of adjacent units

(f) Final batch of adjacent units

Figure C-2. Population Grid with Initial and Follow-up Samples and Areas of Interest. (From EPA QA/G-5S.)

C-14.8. The adaptive SAP design and implementation process for guiding removal actions follows these steps.

C-14.8.1. Sampling location decision points forming a regular grid are laid across the site. Each sample decision point is so named because at each sampling location, the following decision must be made: will this point be removed or left in place? For instance, if the petroleum hydrocarbon concentration at this location exceeds an action level, it will be excavated from the site. An action level serves as the criterion for differentiating among decision points that can be considered clean and points that must be treated as contaminated. Because the acceptable level of uncertainty is very important to the design of the adaptive SAP, it must be determined prior to sampling or before the program begins (i.e., during the data quality objective development process), with mutual agreement from all the stakeholders involved with the site.

C-14.8.2. Based on professional judgment and historical information available for the site, a probability is initially assigned to each decision point; namely, the likelihood contamination at that location is greater than some action level.

C-14.8.3. As sample results become available, the probabilities for each of the decision points are updated with actual data. The site is then divided into three regions: i) the portion of the site (decision points) where the probability that contamination exceeds the action level is low (this region is accepted as clean with perhaps only minimal confirmatory sampling); ii) the portion of the site where the probability of contamination is so high that confirmatory sampling is unnecessary; and iii) the portion of the site where there is neither a high nor low probability of contamination above the action level, i.e., the gray area where there is significant uncertainty whether the presence or absence of contamination is greater than the predetermined action level. Indicator kriging (Appendix R) may be a powerful tool for such an application.

C-14.8.4. Predetermined decision rules are applied. There may be several alternative decision rules that can be used to drive the sampling process. Additional sampling may need to be done for the gray areas, especially if the removal action is desired to lower overall site risk. The decision rules should tend to produce a sampling program that works its way around suspected areas of contamination. The decision rules should also tend to produce a sampling pattern that starts from areas of suspected contamination and works its way outward to the boundary where removal can cease.

C-14.9. Regardless of the decision rule used, the process is the same. Sampling locations are selected that have the greatest opportunity to provide the most benefit in the context of the selected decision rule. After results are obtained, the extent of contamination is re-estimated along with the number of uncertain decision points remaining, and a decision is made where additional removal is justified until no such locations remain.

C-14.10. Figure C-3 shows the adaptive sampling plan process, and Paragraph C-15 illustrates a practical application of an adaptive SAP.

C-15. Case Study 5–Argonne's Adaptive Sampling and Analysis Program. The U.S. Department of Energy's (DOE's) Argonne National Laboratory developed the following case study.

C-15.1. Oil and gas producers may save millions of dollars in cleaning up soils contaminated with naturally occurring radioactive materials by applying an on-site soil sampling and analysis method developed by the U.S. DOE's Argonne National Laboratory.

C-15.2. Naturally occurring radioactive material accumulates when the production of oil and natural gas from underground reservoirs transports small quantities of radium to the surface. Over time, the radium—usually radium-226 and, to a lesser extent, radium-228—can concentrate in pipe scale and sludge deposits, which in turn can contaminate soil and equipment.

C-15.3. The traditional approach to cleaning up such sites involves complicated soil sampling techniques and shipping these samples to off-site laboratories for analysis—a time-consuming and costly process. But a recent demonstration has shown that Argonne's adaptive SAP can dramatically reduce the time and money needed to characterize and remediate sites contaminated with naturally occurring radioactive materials. Adaptive SAP combines real time data collection techniques with in-field decision-making for faster and more precise characterization of a site. It was first used successfully for faster and cheaper cleanup of radioactive contamination at DOE sites.

C-15.4. The demonstration was conducted on a 3.5-acre site at Lease Management, Inc., in Mt. Pleasant, Michigan. Pipe salvaged from nearby oil and gas production sites was stacked there prior to being cleaned and reconditioned. Contaminated scale on the outside of the pipes had fallen off during handling and from exposure to the elements. As a result, soils across the pipe yard had varying levels of radium-226 concentrations.

C-15.5. First, scientists walked over the site with a portable global positioning system and a hand-held gamma ray detection device to map surface gross activity levels. The scientists then used a commercial technology called the RadInSoil™ meter to develop a relationship between gross activity values and radium-226 activity concentrations. State guidelines are based on these activity concentrations. With the field data, researchers then used unique Argonne-developed techniques to determine where soil concentrations of contaminants exceeded regulatory standards and would need to be excavated for disposal. To confirm the presence of radium-226, scientists used a tripod-mounted, camera-like device called a High Purity Germanium gamma spectroscopy system that directly measures radium-226 concentrations in surface soils. With use of the results from adaptive SAP, decisions on excavating contaminated soil for disposal can be made immediately. It took 4 days to characterize and remediate the Michigan site.

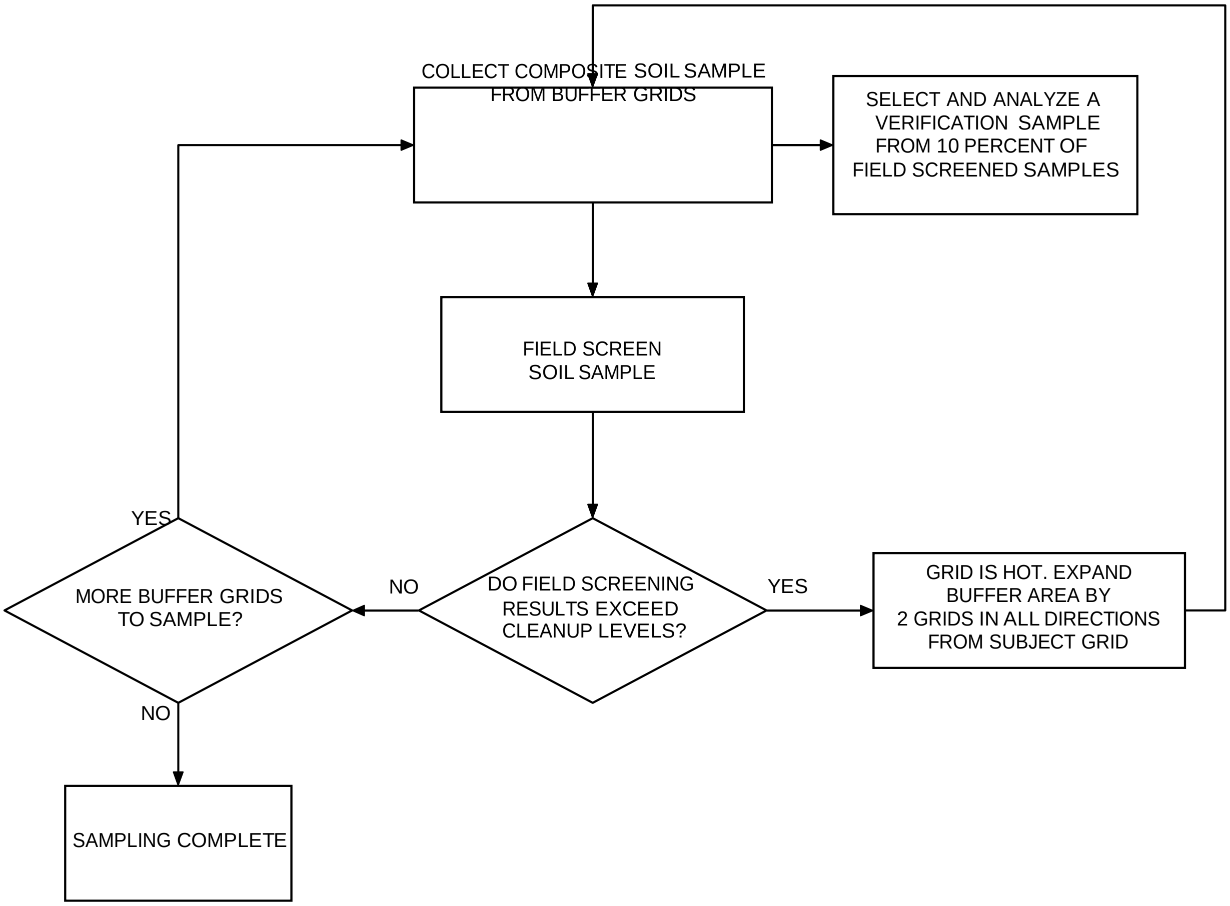

COLLECT COMPOSITE SOIL SAMPLE FROM BUFFER GRIDS
SELECT AND ANALYZE A VERIFICATION SAMPLE FROM 10 PERCENT OF FIELD SCREENED SAMPLES
FIELD SCREEN SOIL SAMPLE
DO FIELD SCREENING RESULTS EXCEED CLEANUP LEVELS?
YES
GRID IS HOT. EXPAND BUFFER AREA BY 2 GRIDS IN ALL DIRECTIONS FROM SUBJECT GRID
NO
MORE BUFFER GRIDS TO SAMPLE?
YES
NO
SAMPLING COMPLETE

Figure C-3. Adaptive Sampling Plan Flow Chart.

C-15.6. The average cost for soil disposal ranges from about $100 to $200 per cubic yard, so keeping soil volumes to an absolute minimum is very important. The goal is to be asprecise as possible in digging up dirt for disposal so one doesn't take anything clean away or leave anything above cleanup standards behind.

C-15.7. For sites contaminated with naturally occurring radioactive materials, it is estimated that using adaptive SAP for site characterization costs only 10% of a more traditional approach. In the Michigan demonstration, the use of adaptive SAP is expected to save the site owner at least $36,000 in disposal costs.

APPENDIX D

Descriptive Statistics

D-1. Introduction. For most environmental sampling, the collected data for some measurement variable of interest constitute a small subset of its set of possible values. The data subset frequently consists of contaminant concentrations from the analysis of environmental (e.g., soil and groundwater) samples collected from the study area. In a statistical context, this subset is referred to as a sample. If it were possible to collect environmental observations from every portion of the study area (i.e., to exhaustively sample an entire site), the set of resulting values would constitute the population. As this is typically not possible, statistics calculated from the sample are used to describe or make inferences about the underlying population. For the environmental applications discussed herein, the statistical methods presented are implicitly for a sample, not the entire population. For more information on populations, the reader is referred to introductory statistical texts readily available in libraries and online.

D-1.1. Commonly used descriptive statistics for environmental data include measures of central tendency, such as mean, median, or mode; measures of relative standing, such as percentiles; measures of dispersion, such as range, variance, standard deviation, coefficient of variation, or interquartile range; measures of distribution symmetry or shape; and measures of association between two or more variables, such as correlation. These measures can also be used to test hypotheses regarding the populations from which the data were drawn.

D-1.2. In general, the sampling design influences how descriptive statistical quantities are calculated. The formulas presented in this monograph are for simple random sampling, simple random sampling with composite samples, and randomized systematic sampling. If more complex designs are used, such as a stratified design, then the formulas need to be adjusted. All of these designs are addressed in Appendix C.

D-1.3. In addition, the distribution of a data set may also influence how descriptive statistical quantities are calculated. Most of the discussion in this Appendix will be centered on normal populations. However, as detailed in Appendix F, it is not uncommon for environmental data to follow other distributions. The most commonly encountered alternative is the lognormal distribution. This Appendix will also present how to calculate the mean and quantiles of the population for a lognormally distributed data set. To estimate other parameters, the reader is urged to refer to any of the excellent texts available, including those referenced here.

D-1.4. The terminology used in presenting general formulas and calculations for this exercise are standard. Out of a total population N, let x_1, x_2, ... , x_n represent the n data points, a sample set of n measurements. Additional information on calculating descriptive statistics for environmental applications can be found in the EPA/240/B-026/003, QA/G-9S and Gilbert (1987).

D-2. Measures of Central Tendency. Measures of central tendency characterize the center of a set of measured data values. The three most common estimates are the mean, median, and mode. These are described below, and examples of calculating each of them are presented in Paragraph D-2.2

D-2.1. Mean. The mean is the most commonly used measure of central tendency. The formula used to calculate the sample mean is a function of the sampling design. The sample mean $\bar{x}$ (arithmetic average) is the sum of the data points, $x_1, x_2, \ldots, x_n$, divided by the total number of data points (n):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{(D-1)}$$

where x_i denotes the value of the i^{th} point.

D-2.1.1. If distribution testing suggests that data are lognormally distributed, then the descriptive statistics are best calculated using the transformed data (for each value $y_i = Ln(x_i)$). Calculating the sample mean, $\bar{x}$, is possible, even for lognormally distributed data. Gilbert (1987) reports that $\bar{x}$ may be used when the population coefficient of variation is small (i.e., less than 1.2). Unfortunately, the sample mean is statistically biased for known lognormal conditions. It is highly sensitive to a few large data values, as is typical of lognormal data. There are alternatives for estimating the population mean that are not statistically biased, and these are preferred.

D-2.1.2. The preferred method for estimating the population mean of a lognormal population is calculated by:

$$\hat{\mu}_1 = e^{\bar{y}} \Psi_n(t) \qquad \text{(D-2)}$$

where

$\bar{y}$ = sample mean of the log-transformed data

n = number of data points

s_y = sample standard deviation of the log-transformed data

$\Psi_n(t)$ (with $t = s_y^2/2$) = the following infinite series

$$\Psi_n(t) = 1 + \frac{(n-1)t}{n} + \frac{(n-1)^3 t^2}{2!\, n^2 (n+1)} + \frac{(n-1)^5 t^3}{3!\, n^3 (n+1)(n+3)} + \frac{(n-1)^7 t^4}{4!\, n^4 (n+1)(n+3)(n+5)} + \ldots$$

D-2.1.3. This is the minimum variance unbiased estimate of the population mean. Likewise, the unbiased estimator of the variance of the mean is:

$$s^2(\hat{\mu}_1) = \exp(2\bar{y})\left\{\left[\Psi_n(t)\right]^2 - \Psi_n\left[t'\right]\right\} \tag{D-3}$$

where

$$t' = \frac{s_y^2(n-2)}{n-1}$$

D-2.1.4. The infinite series may be evaluated on a computer or estimated from tables referenced in Gilbert (1987). This method produces the minimum unbiased variance estimator (statistically unbiased and smallest sampling error variance) of the mean for a lognormal population.

D-2.1.5. Performing this calculation obviously can be laborious. There is a simpler method for estimating the mean and variance of a lognormal population that arises in Gilbert and in EPA guidance documentation. This method uses the formulas:

$$\hat{\mu} = \exp\left(\bar{y} + \frac{s_y^2}{2}\right)$$

$$\hat{\sigma}^2 = \hat{\mu}^2\left[\exp\left(s_y^2\right) - 1\right] \tag{D-4}$$

D-2.1.6. However, the approach can produce poor (biased high) estimates of mean and variance for small data sets and is not recommended unless *n* is large (e.g., *n* > 50). Paragraph D-2.2 presents an example calculation for the mean of a lognormal population using the three methods.

D-2.1.7. For complex sampling designs, such as stratification, the sample mean is a weighted arithmetic average of the sample means of the *L* strata. Because a stratified sampling plan weights the number of samples unequally among areas, the weights for each area are incorporated into the calculation of the average. A weighted average is very similar to the arithmetic average, where an arithmetic average weights each sample result equally (with a weight of 1/*n*). A weighted arithmetic average is calculated by:

$$\bar{x} = \sum_{i=1}^{L} w_i \bar{x}_i \tag{D-5}$$

where:

w_i = weight for the i^{th} stratum

$\bar{x}_i$ = sample mean of the i^{th} stratum

L = number of strata

$$\sum_{i=1}^{L} w_i = 1$$

D-2.1.8. For example, consider a stratified sampling plan that collects a total of n = 20 samples from a site with L = 2 sub-groups, where 8 samples, x_{1i} i = 1,…8, are collected in subgroup 1, and 12 samples, x_{2i} i = 1,…12, are collected in subgroup 2. If the average for the site is required and the two strata are assumed to be of equal area or volume, then the weights for the weighted average are ½ for the sample mean from subgroup 1 and ½ for the sample mean from subgroup 2 so that

$$\sum_{i=1}^{L} w_i = \frac{1}{2} + \frac{1}{2} = 1$$

and the overall mean is

$$\bar{x} = \sum_{i=1}^{L} w_i \bar{x}_i = \frac{1}{2}\bar{x}_1 + \frac{1}{2}\bar{x}_2 = \frac{1}{2}\frac{\sum_{i=1}^{8} x_{1i}}{8} + \frac{1}{2}\frac{\sum_{i=1}^{12} x_{2i}}{12}.$$

D-2.1.9. Careful examination will show that each observation in subgroup 1 is weighted by 1/16 in the overall mean and each observation in subgroup 2 is weighted by 1/24 in the overall mean.

D-2.1.10. The mean is the “center of gravity.” The mean is very sensitive to extreme values because each measurement, x_i, is used to calculate the mean. Note that the sample mean, $\bar{x}$, is distinguished from the corresponding population parameter, the population mean, μ. The population mean could hypothetically be calculated using Equation D-1 if it were possible to exhaustively sample the entire population. The number of all possible data points from the population, N, would appear in the denominator of Equation D-1. Typically, the number of data points in the sample data set, $n << N$ and the sample mean, $\bar{x}$, is a “best” estimate of μ. As previously stated, this section of the document focuses on sample statistics that are ultimately used to estimate the corresponding parameters.

D-2.2. Example of Lognormal Mean Calculations. A group of arsenic measurements in soil were found to be lognormally distributed. The sample analytical results (in mg/kg) are:

SB1	SB2	SB3	SB4	SB5	SB6	SB7	SB8	SB9	SB10
12.461	13.451	13.056	11.502	10.835	30.06	17.72	17.11	12.02	13.73

D-2.2.1. Method 1. Using the simple (albeit biased) population average method, the sample mean of these data is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 15.19 \text{ mg/kg arsenic in soil.}$$

The sample variance $s^2 = 32.3$. Shapiro-Wilk testing (Appendix F) suggests that the lognormal distribution cannot be rejected. Also, the sample variance is high. These data would be better treated as lognormal.

D-2.2.2. Method 2. To calculate the minimum unbiased variance estimator of the mean, we first take the natural logarithm of the data set and calculate the following:

$$\bar{y} = 2.674, \quad s_y^2 = 0.09060, \quad t = \frac{s_y^2}{2} = 0.0453.$$

Using the minimum unbiased variance estimator, we see that the mean is 15.17 mg/kg. Method 1 above, which does not account for the lognormality, is biased high slightly.

D-2.2.3. Method 3. Others may choose to use the simpler Gilbert/EPA estimating method described above. This alternative also yields a sample mean of about 15.17 mg/kg. This result is low relative to the simple averaging method, but in this case is nearly identical to the minimum unbiased variance estimator. This is largely attributable to the low value of *t* in this example.

D-2.2.4. Summary. Ideally, with a computer, the method for minimum unbiased variance estimator of the mean for a lognormal population could be used. In cases of large *n*, it is suitable to use the third, and relatively simpler, method.

D-2.3. Median. The sample median ($\tilde{x}$) is the second most common measure of central tendency. When measurements are ranked from lowest to highest, the median is the middle of the data set. Half of the data are less than the sample median, and half of the data are greater than the sample median.

D-2.3.1. To compute the sample median, 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 so on, where $x_{(n)}$ is the largest.

D-2.3.2. The determination of the sample median depends upon whether the sample size *n* is odd or even:

$$\tilde{x} = \begin{cases} x_{[(n+1)/2]}, & n = 1,3,5.... \\ \dfrac{x_{(n/2)} + x_{[(n/2)+1]}}{2}, & n = 2,4,6... \end{cases}$$

D-2.3.3. The median is also referred to as the 50th percentile, the value greater than or equal to 50 percent of the measurements. Unlike the mean, the median is not influenced by extreme values. The median is also more robust than the mean for censored data (when non-detected results occur). When data are symmetrical, the mean and median of the data are very similar. If data are slightly skewed to higher values, the mean tends to be larger then the median because the mean is more influenced by these higher values than the median. Likewise, when data are skewed to lower values, the mean tends to be lower than the median.

D-2.4. Mode. The third method of measuring the center of the data is the mode. The mode is the value of the sample that occurs with the greatest frequency. To find the mode, count the number of times each value occurs. As this value may not always exist, or if it does, it may not be unique, mode is the least commonly used measure of central tendency; however, it is useful for qualitative data.

D-2.5. Examples for Calculating the Measures of Central Tendency. Consider estimating the measures of central tendency for the subsurface soil background chromium results (in mg/kg) as follows: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, and 3.84.

D-2.5.1. Sample Mean. The sample mean (in mg/kg) is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{4.60 + 5.29 + 4.26 + 5.28 + 4.53 + 5.74 + 5.86 + 3.84}{8} = 4.93\,.$$

(Note that the mean is reported as three significant figures to reflect the minimum number of significant figures in the original data set.)

D-2.5.2. Sample Median. The data, from smallest to largest, are:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} = 3.84, 4.26, 4.53, 4.60, 5.28, 5.29, 5.74, 5.86.$$

As there are eight points (*n* is even), the median (in mg/kg) is:

$$\tilde{x} = \frac{x_{(n/2)} + x_{[(n/2)+1]}}{2} = \frac{x_{(4)} + x_{(5)}}{2} = \frac{4.60 + 5.28}{2} = 4.94.$$

D-2.5.3. Sample Mode. In this example, mode does not exist since no value is repeated multiple times.

D-3. 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 or quantiles are one such measure of relative standing that may also be useful for summarizing data.

a. The percentile is the data value that is greater than or equal to a given percentage of the data values.

b. The quantile is an alternative name for percentile when speaking in fractions (proportions) rather than in percents.

D-3.1. Just as the mean is a measure of location at the center of data, percentiles and quantiles are measures of location at various positions of the data. For a continuous variable X, the $p100^{\text{th}}$ percentile or p quantile, x_p, is the data point that is greater than or equal to $100p\%$ of the data points and is less than or equal to $(1 - p)100\%$ of the data points. For example, if x is the 95% percentile (0.95 quantile), then it has the property that 95% (a proportion 0.95) of the observations lie at or below x_p and 5% (a proportion 0.05) of the data points lie at or above x_p.

D-3.2. The percentile and quantile for a discrete variable (i.e., a variable that may assume only a finite number of values) is defined somewhat differently than for a continuous variable. For a discrete variable X, X_p is the p quantile of X if

$$P(X < X_p) \leq p$$

and

$$P(X > X_p) \leq 1 - p$$

or equivalently,

$P(X \le X_p) \ge p.$

D-3.3. To calculate percentiles or quantiles for a set of n sample points (x_1, x_2, ..., x_n), first list the data points from smallest to largest (x_1, x_2, ..., x_n). Multiply the sample size, n, by p. Divide the result into the integer part and the fractional part, i.e., let $np = j + g$ where j is the integer part and g is the fraction part. The $p100^{th}$ percentile, x_p, is calculated by:

$$x_p = \begin{cases} \dfrac{x_{(j)} + x_{(j+1)}}{2}, & g = 0 \\ x_{(j+1)}, & g \ne 0 \end{cases}$$

D-3.4. One example of a percentile is the median. The median is the 50th percentile because half the results fall below this value and half of the results fall above this value. 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.

D-3.5. Important percentiles usually reviewed are the quartiles of the data. The most common quartiles are 25th, 50th, and 75th percentiles. The 25th and 75th percentiles can be used to estimate the dispersion of a data set (see Paragraph D-4). Quartiles are discussed further in Paragraph D-4 to explain the dispersion of the data.

D-3.6. Also important for environmental data are the 90th, 95th, and 99th percentiles, where a decision-maker would like to be sure that 90, 95, or 99% of the contamination levels are below a fixed risk level. Directions and examples for calculating the measures of relative standing are presented below in Paragraph D-4.

D-3.7. Estimating quantiles in lognormal populations arises frequently in environmental applications. Of course, a probability plot may be used to estimate the quantiles, after the data are transformed and plotted. Alternatively, a mathematical method is recommended in Gilbert (1987). Simply,

$$\hat{x}_p = \exp\left(\bar{y} + Z_p s_y\right) \tag{D-6}$$

where Z_p is the value of the cumulative normal distribution for the pth quantile. For the data in the preceding example (Paragraph D-2.2), the 99th quantile of the data is

$$\hat{x}_{0.95} = \exp\left(2.67 + 2.326 \times 0.301\right) = 29.1 \text{ mg/kg.}$$

D-4. Calculating the Measures of Relative Standing (Percentiles). The 95th, 75th, and 25th percentiles will be computed for the eight subsurface soil background chromium results (in mg/kg), ordered from lowest to highest, as follows: 3.84, 4.26, 4.53, 4.60, 5.28, 5.29, 5.74, and 5.86.

D-4.1. For the 95th percentile, $p = 0.95$ and
$np = (8)(0.95) = 7.6$

Therefore:

$$np = j + g$$

$$7.6 = 7 + 0.6$$

So: $j = 7$ and $g = 0.6$.

D-4.2. Since $g \neq 0$, $x_{(p)} = x_{(j+1)}$. The 95th percentile of this data set is:

$$x_{0.95} = x_{(7+1)} = x_{(8)} = 5.86 \text{ mg/kg}$$

Note that 100% of the data points (8 out of 8 values) rather than 95% of the measurements are less than or equal to the 95th percentile. The 95th percentile is being calculated for the set of eight measured chromium values and not for the set of all possible values of chromium. The set of measured chromium concentrations is a discrete variable (there are only eight possible values for chromium). If a larger number of measurements were made, nearly (or precisely) 95% of the measurements would be less than or equal to the 95th percentile.

D-4.3. For the 75th percentile, $p = 0.75$ and

$$np = (8)(0.75) = 6.$$

Therefore:

$$np = j + g$$

$$6 = 6 + 0.0$$

So: $j = 6$ and $g = 0$.

D-4.4. The 75th percentile of these data is:

$$x_{0.75} = \frac{x_{(6)} + x_{(7)}}{2} = \frac{5.29 + 5.74}{2} = 5.52\,\text{mg/kg.}$$

Note that 6 out of the 8 measured values (0.75 of the total number of observations) are less than or equal to the 75th percentile 5.52 mg/kg.

D-4.5. For the 25th percentile, $p = 0.25$ and

$$np = (8)(0.25) = 2$$

Therefore:

$$np = j + g$$

$$2 = 2 + 0.0$$

So: j = 2 and g = 0.

D-4.6. The 25th percentile of these data is:

$$x_{0.25} = \frac{x_{(2)} + x_{(3)}}{2} = \frac{4.26 + 4.53}{2} = 4.40\,\text{mg/kg.}$$

D-5. Measures of Dispersion.

D-5.1. Introduction. Measures of central tendency are more meaningful if accompanied by information on how the data spread out from the center. Measures of dispersion or variability in a data set include the sample range, variance, standard deviation, coefficient of variation, and the interquartile range. Directions for calculating these measures of dispersion follow, and examples are presented in Paragraph D-6.

D-5.1.1. Range. This is the difference between the largest and smallest result from the data set.

D-5.1.2. Variance. This is a measurement of the dispersion or deviation of results from the mean of a data set.

D-5.1.3. Standard Deviation. This is the square root of the sample variance, it has the same unit of measure as the original data.

D-5.1.4. Coefficient of Variation (CV). This is sometimes called the relative standard deviation (RSD), a unitless measure equal to the standard deviation divided by the mean.

D-5.1.5. Interquartile Range. This is the difference between the 75th and 25th percentiles, it measures the central 50% of the results in the data set.

D-5.2. Sample Range. The simplest measure of dispersion to compute is the sample range. The sample range (R) is the difference between the largest value and the smallest value of the sample:

$$R = x_{(n)} - x_{(1)} \tag{D-7}$$

where:

$x_{(n)}$ = largest ordered value
$x_{(1)}$ = smallest ordered value

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 is greatly influenced by) extreme values.

D-5.3. Sample Variance. The sample variance measures the dispersion or deviation of results from the mean of a data set.

D-5.3.1. To find the sample variance (s^2), compute:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{D-8}$$

D-5.3.2. If the variance is being manually calculated, a simpler version of this calculation is the following:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1} \tag{D-9}$$

D-5.3.3. However, this version should not be used when calculating the variance with a computer because too much rounding error is introduced into this calculation.

D-5.3.4. A large sample variance implies that there is a large spread among the data, that the data are not clustered tightly around the mean. A small sample variance implies that there is little spread among the data, and that most of the data are near the mean. Like the mean, the sample variance is affected by extreme values and by a large number of non-detected results. Note that the sample variance s^2 is distinguished from the corresponding population parameter, the population variance, σ^2.

D-5.4. Sample Standard Deviation. The sample standard deviation has the same unit of measure as the original data. The sample standard deviation (s) is the square root of the sample variance:

$$s = \sqrt{s^2} \tag{D-10}$$

Frequently, the sample standard deviation will not be an appropriate measure of dispersion unless the data are normally distributed.

D-5.5. Sample Coefficient of Variation. The CV or RSD is a unitless measure that

allows the comparison of dispersion across several sets of data because it is scaled to the mean. The sample CV is the sample standard deviation divided by the sample mean:

$$\text{CV} = \frac{s}{\bar{x}} \tag{D-11}$$

The CV is often expressed as a percentage:

$$\%\text{RSD} = \frac{s}{\bar{x}} 100\% \,.$$

The CV is often used in environmental applications because variability (expressed as a standard deviation) is often proportional to the mean.

D-5.6. Sample Interquartile Range (IQR). When extreme values are present, the interquartile range may be more representative of dispersion in the data than the standard deviation. This range is not heavily influenced by extreme values because it measures the spread within the center portion of a data set, rather than include the most extreme values as does the range. As a result, it is useful when the data include a large number of non-detects. Use the directions in Paragraph D-6 to compute the 25th and 75th percentiles of the data ($x_{0.25}$ and $x_{0.75}$ respectively). Then,

$$\text{IQR} = x_{0.75} - x_{0.25} \tag{D-12}$$

D-6. Examples for Calculating the Measures of Dispersion. Consider estimating the measures of dispersion for subsurface soil chromium results (in mg/kg) as follows: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, and 3.84. The data are ordered as follows:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} = 3.84\,, 4.26, 4.53\,, 4.60, 5.28, 5.29, 5.74, 5.86\,.$$

D-6.1. Sample Range (R). The sample range is simply:

$$R = x_{(n)} - x_{(1)} = 5.86 - 3.84 = 2.02$$

D-6.2. Sample Variance (s^2). Before the variance can be computed, the mean must be computed. The mean was computed in Paragraph D-2.2 and is 4.93 mg/kg. Both methods of calculating the variance are illustrated below:

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

$$= \frac{(4.60-4.925)^2 + (5.29-4.925)^2 + (4.26-4.925)^2 + (5.28-4.925)^2}{8-1} + \frac{(4.53-4.925)^2 + (5.74-4.925)^2 + (5.86-4.925)^2 + (3.84-4.925)^2}{8-1}$$

$$= 0.5255$$

or

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

$$= \frac{\left(4.60^2 + 5.29^2 + 4.26^2 + 5.28^2 + 4.53^2 + 5.74^2 + 5.86^2 + 3.84^2\right)}{8-1} - \frac{\left(8 \times 4.925^2\right)}{8-1}$$

$$= 0.5255$$

D-6.3. Sample Standard Deviation (s).

$$s = \sqrt{s^2} = \sqrt{0.5255} = 0.7249$$

D-6.4. Sample Coefficient of Variation (CV).

$$\mathrm{CV} = \frac{s}{\bar{x}} = \frac{0.7249}{4.925} = 0.1472$$

D-6.5. Sample Interquartile Range (IQR). The 25th and 75th percentiles of the data, $x_{0.25}$ and $x_{0.75}$ respectively, were computed in Paragraph D-4. So:

$$\mathrm{IQR} = x_{0.75} - x_{0.25} = 5.515 - 4.395 = 1.12$$

Note that the single data set presented above results in a number of different numerical values that all summarize dispersion:

Range	IQR	s	s^2	CV
2.0 mg/kg	1.1 mg/kg	0.72 mg/kg	0.52 mg^2/kg^2	0.15

APPENDIX E

Statistical Distributions

Section I
Introduction

E-1. One of the essential decisions that precedes many statistical calculations is determining the statistical distribution. Whether the data can be classified as normally distributed, lognormally distributed, meeting some other distribution, or meeting no distributional assumption, dictates how subsequent calculations and statistical tests are chosen and conduced. Distributional assumptions are common in statistical analyses, especially assumptions of normality. Data from environmental studies tend to be skewed rather than following a classical bell-shaped curve, or normal distribution. Thus, verifying distributional assumptions is critical to a successful statistical analysis.

E-2. To provide an objective basis for making this decision, statistical tests are available and discussed in this Appendix. Tests can be applied to the untransformed data when testing for normality or to the log-transformed data when testing for lognormality. Normal probability plots should also be constructed and examined as described in Appendix J.

Section II
Probability Distributions

E-3. Introduction. Many statistical tests and models are appropriate only for data that follow a particular distribution. For a continuous variable X (e.g., the concentration of a contaminant), the distribution is modeled by a mathematical function of the form: $P = P(X)$, where $P(X)$ is referred to as the probability density function or probability distribution. A plot of P versus X generates a curve. The area (integral) under the curve between any two points, X_a and X_b, gives the probability that the random variable X lies between the two points, $P(X_a \leq X \leq X_b)$, which will be a number between 0 and 1. The total area under the entire curve is always 1. Figure E-1 plots $P(X)$ and shows how $P(5 < X < 6)$ would be found.

E-3.1. A common use of probability density functions is to calculate population percentiles for the distribution. For example, if $X_{0.95}$ is the value such that $P(X \leq X_{0.95}) = 0.95$, then $X_{0.95}$ is referred to as the 95th (population) percentile or 0.95 quantile of X. In general, X_p denotes the $p100^{th}$ percentile or p quantile of X. Appendix D covers techniques to estimate the population percentile from sample data.

E-3.2. Two of the most important distributions for tests involving environmental data are the normal and the lognormal probability distributions. When a parametric statistical test is performed on some set of measured values of X ($x_1, x_2, \ldots, x_n$), some specific probability density function, $P(X)$, is either known or assumed. This section will provide guidance for determining if the distributional assumption of a given statistical test is satisfied; in particular, the assumption of normality, as this assumption is fundamental to virtually all parametric statistical tests.

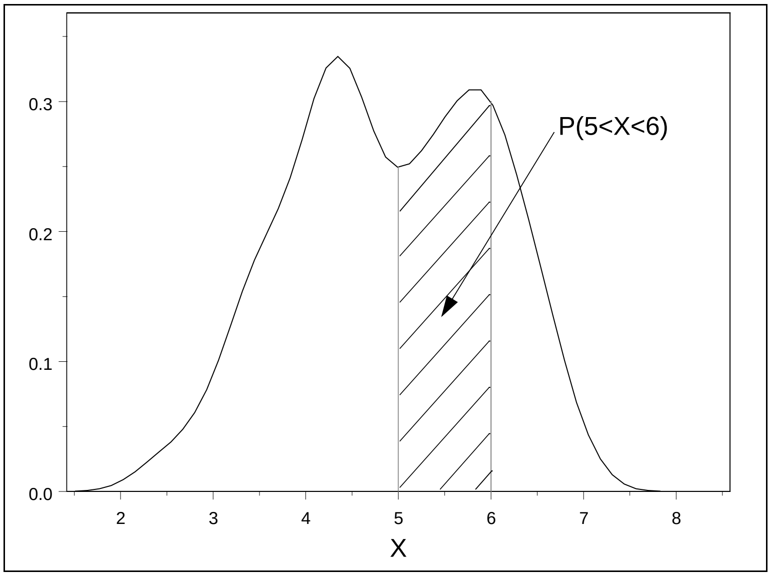


Figure E-1. Probability Density Function.

E-3.3. Normal Distribution. If the variable *X* possesses a normal or Gaussian distribution (i.e., is said to be normally distributed), then the probability density function for *X* is

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2} \right) . \qquad \text{(E-1)}$$

E-3.3.1. A plot of *X* versus *P*(*X*) generates a bell-shaped curve. Two such curves are shown in Figure E-2. The function *P*(*X*) depends on two parameters (constants), the population mean, μ, and the population standard deviation, σ, where $\sigma > 0$. It is often useful to work with the square of the standard deviation, σ^2, which is referred to as the population variance. Note that the normal distribution is symmetrically centered about the mean, μ, and tapers off rapidly at the tails.

Because exactly 50% of the distribution falls below the mean, the median (50th percentile) of the normal distribution is equal to the mean. The value of the parameter σ affects the shape of the distribution. In particular, as shown in Figure E-2, as the value of the standard deviation is increased from σ_1 to some value $\sigma_2 > \sigma_1$, the "spread" of the distribution about the mean increases. Because a normal distribution depends upon the parameters, μ and σ, it is often denoted by $N(\mu,\sigma)$.

E-3.3.2. The normal distribution is critical because measurement data (e.g., a set of concentration measurements) can often be modeled by it. When it is known or it can be assumed that a set of measurements, $x_1, x_2, \ldots, x_n$, follow a normal distribution, then the sample mean, $\bar{x}$, is a good estimate of the population mean, μ. Also the sample standard deviation, s, is a good estimate for the population standard deviation, σ. (Refer to Appendix D for the definitions of the sample mean and standard deviation.)

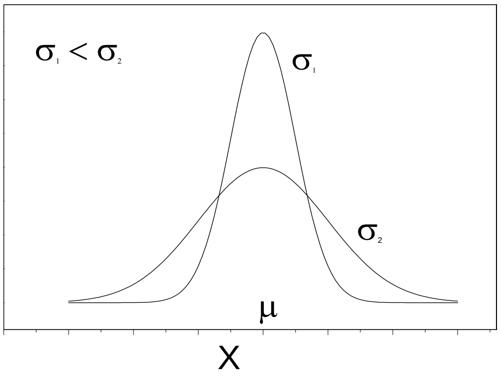


Figure E-2. Normal Distribution.

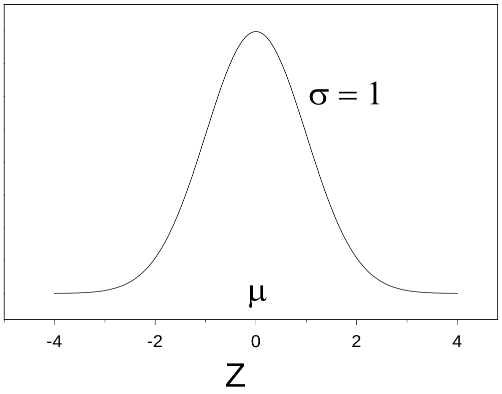


Figure E-3. Standard Normal (*Z*) Curve.

E-3.3.3. It can be shown, if the random variable *X* possesses a normal distribution, then the random variable has a standard normal distribution, $N(0,1)$.

$$Z = \frac{(X - \mu)}{\sigma} \qquad \text{(E-2)}$$

The probability density function of the standard normal distribution is illustrated in Figure E-3. Using the notation from above, we can denote the *p*100th percentile (*p* quantile) of Z as Z_p. The standard normal distribution is important since the percentiles Z_p are commonly listed in statistical tables like Table B-15.

E-3.3.4. For example, if random variable *X* is $N(3,2)$, we can use Table B-15 to find $X_{0.95}$ as follows. Find the closest value to 0.95 in the interior of Table B-15. In this case 0.9495 and 0.9505 are equally distant. Find $Z_{0.95}$ by the value to the far left of the row found in the last step and the top of the column. Here, it is necessary to interpolate between 1.64 and 1.65 to get $Z_{0.95} = 1.645$. Figure E-4 demonstrates that 95% of the area under the standard normal density curve (the shaded area) lies to the left of 1.645. Returning to the stated problem, solve Equation E-2 for *X* to get:

$$X_p = \mu + Z_p \sigma \qquad \text{(E-3)}$$

so in this example,

$$X_{0.95} = 3 + 1.645(2) = 6.29 .$$

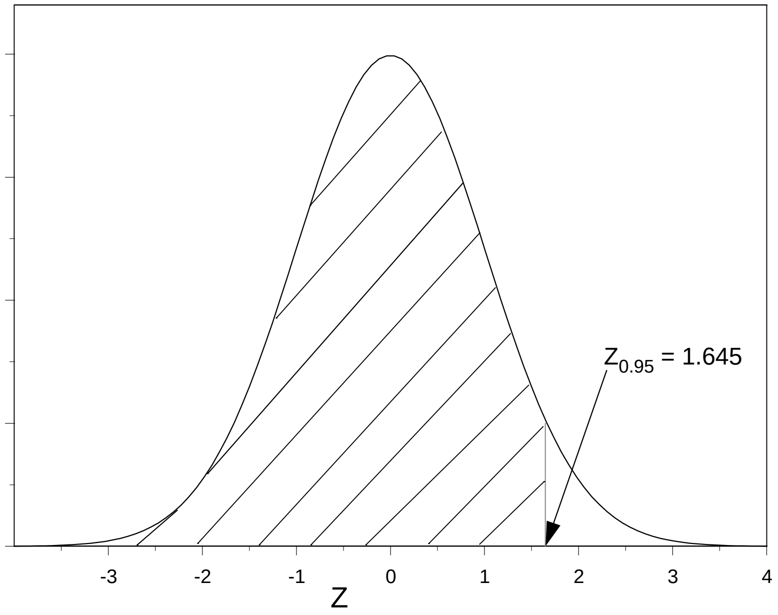


Figure E-4. 95th Percentile of the Standard Normal Distribution.

E-3.3.5. Because the standard normal distribution is symmetrical about a mean of zero, $Z_{1-\alpha} = -Z_{\alpha}$. Thus, the area of the standard normal curve that falls between $Z_{1-\alpha}$ and Z_{α} is equal to $1 - 2\alpha$ (e.g., for $\alpha = 0.05$, 90% of the distribution falls between $Z_{0.05} = -1.645$ to $Z_{0.95} = 1.645$). It follows from Equation E-1 that, in terms of the variable X, the proportion $1 - 2\alpha$ (equivalently, $100(1 - 2\alpha)\%$) of the distribution falls between $X_{\alpha} = \mu + Z_{\alpha}\sigma$ and $X_{1\text{-}\alpha} = \mu + Z_{1-\alpha}\sigma$. Because $Z_{1-\alpha} = -Z_{\alpha}$, $100(1 - 2\alpha)\%$ of the distribution falls within $\mu \pm Z_{1-\alpha}\sigma$. Some examples are presented below:

a.90% of the distribution ($\alpha = 0.05$) falls within the interval $\mu \pm 1.645\sigma$.

b.95% of the distribution ($\alpha = 0.025$) falls within the interval $\mu \pm 1.960\sigma$.

c.99% of the distribution ($\alpha = 0.005$) falls within the interval $\mu \pm 2.576\sigma$.

d.99.9% of the distribution ($\alpha = 0.0005$) falls within the interval $\mu \pm 3.291\sigma$.

E-3.3.6. Thus, approximately 95% of the distribution falls within two standard deviations of the mean ($\mu \pm 2\sigma$) and over 99% (in fact, about 99.7%) of the distribution falls within three standard deviations of the mean ($\mu \pm 3\sigma$). It can similarly be shown that about 68% of the distribution falls within one standard deviation of the mean.

E-3.3.7. Finally, a useful property of the normal distributions is that that any linear combination of normally distributed variables will also be normally distributed. In particular, let

$$Y = \frac{\left(X_1 + X_2 + \cdots + X_n\right)}{n}$$

where each random variable X_i follows the same normal distribution $N(\mu, \sigma)$. It can be shown that the random variable *Y* is distributed as

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

This is extremely useful because the definition of *Y* is very similar to the definition of the sample mean, $\bar{x}$, presented in Appendix G. Thus, if the variable *X* is normally distributed, with mean μ and standard deviation σ, a set of *n* measurements of *X* are taken, the sample mean $\bar{x}$ is calculated for the set of *n* measurements, and this process could be repeated indefinitely. The resulting distribution of values of the sample mean will be normally distributed with mean and standard deviation:

$$\mu_{\bar{x}} = \mu\,, \quad \sigma_{\bar{x}} = \sigma / \sqrt{n}$$

E-3.3.8. It also follows that

$$Z = \frac{(\bar{x} - \mu)}{\left(\sigma / \sqrt{n}\right)} \tag{E-4}$$

will follow a standard normal distribution. Although σ is not typically known, it can be shown that for sufficiently large *n*, is closely approximated by a standard normal

$$Z = \frac{(\bar{x} - \mu)}{\left(s / \sqrt{n}\right)} \tag{E-5}$$

distribution. Furthermore, if *X* is normally distributed and *n* is large, then an approximate *p*100% upper bound can be calculated for the population mean from the above equation.

$$\mu \le \bar{x} + Z_p (s / \sqrt{n}) \tag{E-6}$$

E-3.3.9. The right side of inequality is approximately the *p100%* upper one-sided confidence limit (UCL) of the population mean. For example, if *p* = 0.95, then the right side of the inequality is the 95% UCL of the population mean. For *p* = 0.95, the population mean μ will be less than the UCL an average of 95 out of 100 times. The calculation of a 95% UCL is typically used in environmental risk assessments.

E-3.3.10. Lastly, it should be noted that the UCL is useful because of the central limit theorem. According to the central limit theorem, Equation E-6 is approximately valid for n sufficiently large regardless of whether or not the measurement variable X is normally distributed. The central limits theorem is discussed below.

E-4. Central Limit Theorem. The central limit theorem states:

> "If a variable X possesses ANY probability distribution with mean (μ) and finite standard deviation (σ), then the sample mean ($\bar{x}$) will be approximately normally distributed with mean (μ) and standard deviation ($\sigma / \sqrt{n}$) if n is sufficiently large."

E-4.1. In other words, if a set of n data points is collected and the sample mean is calculated, and this process is repeated many times and all the resulting values of sample mean are plotted (on a histogram), then the resulting distribution will be approximately normal if n is large (i.e., $n > 50$). As the size of the sample increases, the mean of that sample acts increasingly as if it came from a normal distribution regardless of the true distribution of the individual values. As a consequence, statistical tests that require normality may be performed using the sample mean. Thus, large sample sizes are desirable within the limits imposed by available resources.

E-4.2. The central limit theorem is important for environmental applications, because the mean of a random sample of observations or measurements is frequently of interest (for example, to calculate an exposure point concentration for a risk assessment). Furthermore, no actual environmental data set is completely normal. The assumption of normality for any data set will always be an approximation. In many cases, the normality based statistical tests are not overly affected by a small or even moderate deviation from normality as the tests are robust (sturdy) and perform tolerably well, unless gross non-normality is present. The central limit theorem ensures that tests become increasingly tolerant of deviations from normality as the number of individual samples constituting the sample mean increases.

E-5. Student's *t* Distribution. The Student's *t* distribution is a continuous probability distribution that is similar in shape to the standard normal distribution. Like the standard normal distribution, the *t* distribution is a bell-shaped curve that is symmetrical about a mean of zero. However, the *t* distribution is somewhat flatter in the center and possesses fatter tails than the standard normal distribution. Furthermore, the shape of the *t* distribution is dependent upon the "degrees of freedom," ν (the Greek letter nu). Each value of ν (ν = 1, 2, 3 …) gives rise to a different *t* distribution curve. The degree of "fatness" in the tails of a *t* distribution depends upon the value of ν. As ν increases, the *t* distribution approaches a normal distribution. These properties are illustrated in Figure E-5. For most practical applications, the *t* distribution may be approximated using a standard normal distribution when $\nu > 30$. The mathematical function that defines the probability distribution is more complex than that for the normal distribution and is not presented.

E-5.1. The standard normal curve is used when the mean (μ) and standard deviation (σ) of a normally distributed population of interest are known. When only an estimate of the standard deviation (s) is available from a sample, the t distribution applies. More precisely, if the variable X possesses a normal distribution, then the variable:

$$t_{\nu} = \frac{\bar{x} - \mu}{s / \sqrt{n}} \tag{E-7}$$

possesses a t distribution with $\nu = n - 1$ degrees of freedom. The p100% percentiles (p quantiles) of the t distribution are denoted as $t_{p,\nu}$. This value can be found using Table B-23. Find the row matching the degrees of freedom, ν, on the left side of the table. Find the column containing the value p along the top of the table. The value of $t_{p,\nu}$ is found at the intersection of this row and column. For example, $t_{0.95,10} = 1.812$.

E-5.2. Note that the equation that defines $t_{p,\nu}$ provides the basis for calculating an upper bound for the mean (μ) when μ is unknown but the sample mean is normally distributed. It can be shown that

$$\mu \le \bar{x} + t_{p,\nu} (s / \sqrt{n}) \tag{E-8}$$

where the sample mean ($\bar{x}$) and the sample standard deviation (s) are calculated for some set of n data points and the value $t_{p,\nu}$ is obtained from Table B-23. Roughly speaking, the probability that the population mean will be less than or equal to the right side of the above inequality is p100%. The right side of the above inequality is referred to as the upper one-sided p100% confidence limit of the population mean or simply as the 95% UCL of the population mean.

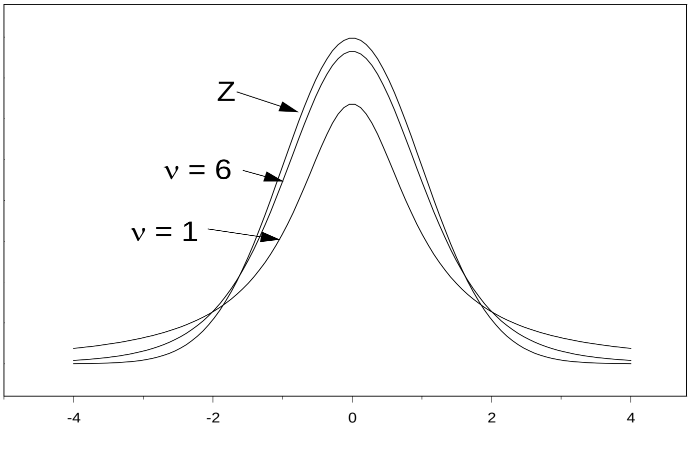


Figure E-5. Comparison of Student's t-Distribution with Standard Normal Distribution.

E-6. Lognormal Distribution. It is not uncommon for environmental data to follow a lognormal distribution. Data collected from contaminated sites often possess a skewed probability distribution that is easily modeled by a lognormal distribution (EPA 600/R-97/006). This occurs because contaminant concentrations are constrained to be non-zero values, with very high values near a source and declining contaminant concentrations away from source areas.

E-6.1. The lognormal distribution is a continuous, non-symmetrical, positively skewed probability distribution that is bounded to the left by zero. However, like the normal distribution, the lognormal distribution is completely characterized by two parameters that represent the population mean and standard deviation of the log-transformed distribution. Several lognormal distributions are shown in Figure E-6.

E-6.2. There is a simple relationship between the normal and lognormal distributions. If X is lognormally distributed, then $Y = Ln(X)$ is normally distributed. Though the probability distribution is a non-symmetrical, positively skewed curve (where the median of the distribution is less than the mean), the probability distribution for $Y = Ln(X)$ is the symmetrical, bell-shaped normal curve. It is a common practice to transform data using the natural log function to achieve approximate normality prior to conducting statistical tests. Just as the notation $N(\mu, \sigma)$ was used to denote a normal distribution, a lognormal distribution will be denoted by $\Lambda(\mu, \sigma^2)$, where μ and σ^2, denote the population mean and variance, respectively, of the normally distributed variable $Y = Ln(X)$ (rather than the lognormally distributed variable X). For brevity, the following notation will be used to indicate that X possesses a log normal distribution: $X \sim \Lambda(\mu, \sigma^2)$, or, equivalently, $Ln(X) \sim N(\mu, \sigma)$.

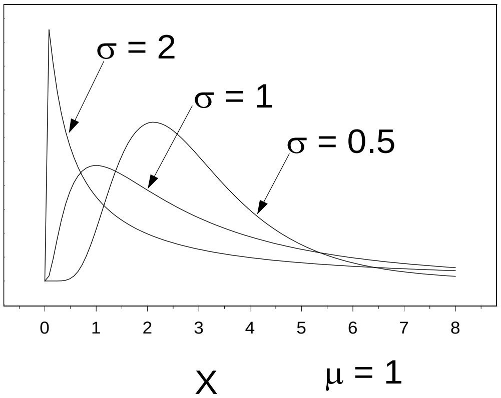


Figure E-6. Lognormal Distributions.

E-6.3. Because any linear combination of normally independent distributed variables will be also be normally distributed, owing to the relationship $Y = Ln(X)$, the product a set of independent lognormally distributed variables will also be lognormally distributed. For example, if $X_1 \sim \Lambda(\mu_1, \sigma_1^2)$, and $X_2 \sim \Lambda(\mu_2, \sigma_2^2)$, then

$$X_1 X_2 \sim \Lambda(\mu_{1+}\,\mu_2,\ \sigma_1^{\,2} + \sigma_2^{\,2})$$

$$X_1/X_2 \sim \Lambda(\mu_1 - \mu_2,\ \sigma_1^{\,2} + \sigma_2^{\,2}).$$

Also, if $X \sim \Lambda(\mu, \sigma^2)$, then

$$cX^b \sim \Lambda(a\mu + b,\ b^2\sigma^2)$$

where c and b are constants, where $c = \exp(a) > 0$ and $b \neq 0$.

E-6.4. The lognormal distribution $\Lambda(\mu, \sigma^2)$ is mathematically described by:

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{X}\exp\left(-\frac{\left(Ln(X)-\mu\right)^2}{2\sigma^2}\right)\ . \tag{E-9}$$

The population mean, μ_X, and standard deviation, σ_X, of the lognormally distributed variable X are calculated as:

$$\mu_X = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{E-10}$$

$$\sigma_X^2 = \exp\left(2\mu + \sigma^2\right)\left[\exp\left(\sigma^2\right) - 1\right] \ = \mu_X^2\,[\exp(\sigma^2) - 1]\,. \tag{E-11}$$

It follows that the (population) coefficient of variation of X is

$$CV = \mu_X \,/\, \sigma_X = \left[\exp(\sigma^2) - 1\right]^{1/2}.$$

The p100% population percentile (p quantile), X_p, can be found from the corresponding p100% percentile of the standard normal distribution, Z_p, as follows:

$$X_P = \exp\left(\mu + Z_p\sigma\right). \tag{E-12}$$

E-7. Binomial Distribution. The binomial distribution is useful in describing the number of successful outcomes, K, from a set number of observations, n. The distribution is considered binomial if the following conditions are satisfied (Moore, 1999):

a. The number of observations, n, is fixed.

b. The n observations are all independent; that is, each observation has no effect on any other.

c. Each observation falls into one of two mutually exclusive categories: Each observation is either a “success” or a “failure.”

d. The probability each observation is a “success” is p. (The probability each observation is a “failure” is $1-p$).

E-7.1. A common example that gives rise to a binomial distribution would be counting the number of heads (successes) obtained from flipping a coin a set number of times. As the number of successful outcomes, K, is a discrete rather than continuous random variable, then the value of the variable K can equal any integer value from 0 to n. The binomial probability distribution is described mathematically by:

$$P(K = k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k} \quad . \tag{E-13}$$

The population mean, μ, and standard deviation, σ, are given by:

$$\mu = np \tag{E-14}$$

$$\sigma = \sqrt{np(1-p)} \quad . \tag{E-15}$$

E-7.2. Table B-1 gives probabilities for the binomial in terms of cumulative probability distribution. That is, the table reports:

$$P(K \le k) = \sum_{i=1}^{k} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \quad . \tag{E-16}$$

For example, for n = 4 and p = 0.5, $P(K \le 2) = 0.6875$.

E-7.3. The binomial distribution under certain conditions can be related to the normal distribution (and the Poisson distribution, as seen in Paragraph E-8). In particular, as n becomes large, the binomial distribution gets close to a normal distribution with mean, np, and standard deviation, $\sqrt{np(1-p)}$. As a rule, this approximation should be used only when both np and $n(1-p)$ are larger than 10 (Moore, 1999).

E-8. Poisson Distribution. The Poisson distribution is useful in describing the number of occurrences of an event over a fixed interval of time. A distribution is considered a Poisson distribution if the following conditions are satisfied:

a. The event is a rare occurrence.

b. The occurrence of two or more events in a small interval of time is zero.

c. A large number of independent observations are made.

d. The average number of occurrences, λ,, over some fixed interval of time is constant (Mason et al., 1989).

E-8.1. The Poisson distribution is typically used to describe or predict rare events. Data from a Poisson distribution must be independent and must be composed of only two responses, such as detected or not detected. Poisson distributions are common when counting the number of detected or not detected occurrences with environmental data that contain only a small percentage of detected concentrations. The probability for one of the two mutually exclusive outcomes must be small. Therefore, the Poisson distribution can be used for highly censored environmental data because the detection of an analyte in a sample would constitute a rare event. This often occurs for background data when organics are being analyzed (most of the results are reported as not detected).

E-8.2. The Poisson distribution can be used with background data to calculate upper limits for the number of detections for each organic analyte. The limits would subsequently be compared to the study area data to determine if detections for a given organic analyte are being obtained more frequently for the study area than for the background
area.

E-8.3. The Poisson distribution may be used for highly censored environmental data in one of two ways. In the first approach, X denotes the number of times an analyte is detected. If the variable X follows a Poisson distribution, then the probability density function is described mathematically by:

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!} \tag{E-17}$$

where μ denotes the mean of the Poisson distribution (such as the average number of times the analyte is detected). For example, if n analyses are performed (n background wells are analyzed for an analyte) and the analyte is detected k times, then the average number of detections, μ, is approximately:

$$\mu \approx \bar{x} = \frac{k}{n}.$$

Data following a Poisson distribution have an equal mean and variance (i.e., $\mu = \sigma^2$).

E-8.4. When n is large and p is small, the binomial distribution and the Poisson distribution give similar results. If follows from Equation E-14 that the probability of detecting the given analyte k out of n times can be calculated using the binomial distribution using the relationship:

$$p = \frac{\mu}{n} \approx \frac{k}{n^2} \ .$$

Therefore,

$$\left\{ P(X = x) = \frac{\mu^x e^{-\mu}}{x!} \right\} \approx \left\{ P(K = k) = \frac{n!}{k!(n-k)!} (\mu / n)^k (1 - (\mu / n))^{n-k} \right\}.$$

E-8.5 For example, if $k = 6$ and $n = 100$, then

$$\mu \approx \bar{x} = \frac{6}{100} = 0.06$$

and

$$p = \frac{\mu}{n} \approx \frac{k}{n^2} = \frac{6}{100^2} = 0.0006.$$

Using the Poisson distribution, we find that the probability of one detection is

$$P(X = 1) = \frac{0.06^1 e^{-0.06}}{1!} = 0.056506 \ .$$

Using the binomial distribution, we find that the probability is:

$$P(K = 1) = \frac{100!}{1!\,(100-1)!} (0.0006)^1 (1 - 0.0006)^{100-1} = 0.056539 \ .$$

As previously stated, these probabilities are very similar as p is small and n is large.

E-8.6. In a second approach, X may denote the concentration per sample rather than the number of detections. In this context, sometimes referred to as the "molecular approach," n samples are analyzed, the analyte is detected in the i^{th} sample at a concentration of x_i, and units for the n measurements are selected such that $x_i > 1$. For example, $x_i = 2\,\mu\text{g} / \text{L} = 2\,\text{ppb}$. In this example, the i^{th} sample is detected at two units or occurrences per billion units of sample examined. (The Poisson distribution is appropriate since the ratio of analyte to sample is small.) The mean concentration per sample (mean number of units per billion units of sample examined) will be:

$$\mu \approx \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}. \tag{E-18}$$

Using this approach, we can readily calculate the probability that the analyte will be detected at a concentration X when X is a whole number.

E-8.7. Note the difference between the two approaches. For the first approach, the mean number of detections for a set of n samples is being calculated. A detection, regardless of the magnitude of the reported concentration greater than the detection limit, consists of a unit count for the calculation of the mean. In the second approach, the mean concentration or number of counts per sample is being calculated; thus, the magnitude of detected concentrations for an individual sample influences the estimation of the mean.

E-8.8. A useful property of the Poisson distribution is that, if the independent variables $X_1, X_2 \ldots X_n$ possess Poisson distributions with means $\mu_1, \mu_2 \ldots \mu_n$, respectively, then the sum of the variables

$$Y = \sum_{i=1}^{n} X_i$$

has a Poisson distribution with mean

$$\mu_Y = \sum_{i=1}^{n} \mu_i \quad .$$

Therefore, if all of the means $\mu_i = \mu$, it follows that $\mu_Y = n\mu$ and $\mu = \frac{\mu_Y}{n} \approx \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$.

E-8.9. As the parameter, μ, becomes very large, the Poisson distribution can also be approximated by a normal distribution. In this case the mean and variance of the normal distribution equal to μ.

E-9. Nonparametric (Distribution Free). Nonparametric statistical methods are used when it is inappropriate to assume some underlying distribution for a data set (when a data set does not conform to some desired theoretical probability distribution). Sometimes it is difficult to verify or satisfy the assumptions that are associated with parametric distributions, such as normal and lognormal distributions for environmental data sets. Using parametric statistical tests when the appropriate assumptions have not been met can result in inaccurate conclusions. In this situation, nonparametric (distribution free) statistical procedures would be appropriate and recommended (Gilbert, 1987; Hahn and Meeker, 1991).

APPENDIX F

Testing for Normality

Section I
Methods for Determining Normality

F-1. Introduction. As previously stated, the assumption of normality is important because it is required for many statistical tests. A normal, or Gaussian, distribution is one of the most common probability distributions used for 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 models follow a normal distribution. Therefore, this Appendix will focus on statistical tests that are used to determine whether normality can be reasonable assumed for a set of measured results.

F-1.1. In general, any distribution assumption should be verified using a combination of graphical plots and statistical tests. Environmental data commonly exhibit frequency distributions that are non-negative and positively skewed (i.e., possess long right tails). Several parametric probability distributions have these properties, including the Weibull, gamma, and lognormal distributions. The methods for testing for normality described in this Appendix can be used to test for lognormality if a logarithmic transformation has been applied to the data.

F-1.2. There are many methods available for verifying the assumption of normality, ranging from simple to complex. They are listed in Table F-1 below. It should be noted that statistical tests for normality do not actually demonstrate normality but the lack of normality. They rely on the probability a given data set is normal (e.g., statistical software typically reports a “*p* value” for the hypothesis that the population distribution is normal). If the probability is low (e.g. $p < 0.01$), one “rejects the assumption of normality,” that is, one concludes, based upon weight of evidence, that the data set is not normal. However, if the assumption of normality is not rejected, then, strictly speaking, the statistical test is inconclusive; the data may or may not be normal. This constitutes an additional reason to visually examine the data set for normality and to decide whether to proceed with a statistical test that requires normality. In practice, if the assumption of normality is not rejected and graphical plots suggest normality, the statistical tests that rely upon normality are typically used.

Table F-1.
Methods Available To Verify the Assumption of Normality

Test	Sample Size, *n*	Recommended Use
Graphical Methods	Any	Highly recommended in conjunction with test methods.
Shapiro-Wilk *W* Test	≤ 50	Highly recommended (D'Agostino's test may be used when sample size is between 50 and 1000).
Filliben's Statistic	≤ 100	Highly recommended.
Coefficient of Variation Test	Any	Only use to quickly discard an assumption of normality and for screening only.
Geary's Test	> 50	Useful when tables for other tests are not available.
Studentized Range Test	≤ 1000	Use for screening purposes only.
Chi-square Test	Large	Useful for grouped data and when the comparison distribution is known.
Lilliefors Kolmogorov-Smirnoff Test	> 50	Useful when tables for other tests are not available.

F-2. Graphical Methods.

F-2.1. Graphical methods present qualitative information about data sets that may not be apparent from statistical tests. Histograms and normal probability plots are some graphical methods that are useful for determining whether data follow a normal curve. The histogram of a normal distribution is bell-shaped. The normal probability plot (Appendix J) 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. Extreme deviations from normality are often readily identified from graphical methods. However, in many instances the decision is not straightforward. Using a plot to decide whether a data set is normally distributed involves making a subjective decision; formal test procedures are usually necessary to test the assumption of normality.

F-2.2. In general, both statistical tests and graphical plots should be used to evaluate normality. The assumption of normality should not be rejected on the basis of a statistical test alone. In particular, when a large number of data are available, statistical tests for normality can be sensitive to very small (i.e., negligible) deviations in normality. Therefore, if a very large number of data are available, a statistical test may reject the assumption of normality when the data set, as shown using graphical methods, is essentially normal and the deviation from normality too small to be of practical significance.

F-3. Shapiro-Wilk Test for Normality.

F-3.1. General. One of the most powerful and most commonly employed tests for normality is the *W* test by Shapiro and Wilk, also called the Shapiro-Wilk test. The Shapiro-Wilk test is an effective method for testing whether a data set has been drawn from an underlying normal distribution. It can also evaluate lognormality if the test is conducted on logarithms of the data. 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 (the data follow a normal curve), the test statistic will be relatively high. If the normal probability plot has curvature that is evidence of non-normality in the tails of a distribution, the test statistic will be relatively low. The Shapiro-Wilk test is recommended in several EPA guidance documents and in many statistical texts. It is designed so that the burden of proof rests on showing evidence that the data are not normally distributed. (In terms of hypothesis testing, the Shapiro-Wilk test is based on H_0 that the data are normally distributed. Hypothesis testing is addressed in detail in Appendices L, M, and N.)

F-3.1.1. The Shapiro-Wilk test is good for evaluating whether a sample set of data has been drawn from a normal or lognormal distribution. However, this test will not have very much power to reject the null hypothesis of normality or lognormality if the sample size is very small (i.e., the test would fail to detect non-normal behavior when the sample size is small). The method for calculating the *W* statistic is presented below in Paragraph F-3.2.

F-3.1.2. As this test is laborious to compute by hand, statistical software packages such as SAS, WQ Stat, and Statistica. An example calculation is presented below in Paragraph F-3.3.

F-3.1.3. D’Agostino’s test is an extension of the Shapiro-Wilk test. It is based on an estimate of the standard deviation obtained using the ranks of the data. This estimate is compared to the usual estimate of the standard deviation, which is appropriate for the normal distribution. The D’Agostino’s test is recommended for sample sizes between 50 and 1000.

F-3.1.4. Another test related to the *W* test is Filliben’s statistic, also called the probability plot correlation coefficient. This statistic measures the linearity of the points on the normal probability plot. Similar to the Shapiro-Wilk test, if the normal probability plot is approximately linear (the data follow a normal curve), the correlation coefficient will be relatively high. If the normal probability plot contains significant curves (the data do not follow a normal curve), the correlation coefficient will be relatively low. Filliben’s statistic is recommended for sample sizes less than or equal to 100. Although easier to compute than the Shapiro-Wilk test, Filliben’s statistic is still difficult to compute by hand. It is available in various software packages.

F-3.2. Directions for the Shapiro-Wilk W Test. Order the data points, $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$, where $x_{(1)}$ is the smallest value and $x_{(n)}$ is the largest value of the *n* observations.

F-3.2.1. Estimate the sample standard deviation, *s*. Compute the Shapiro-Wilk test statistic:

$$W = \left[\frac{b}{s\sqrt{n-1}} \right]^2$$

where

$$b = \sum_{i=1}^{k} b_i = \sum_{i=1}^{k} a_{n-i+1} [x_{(n-i+1)} - x_i].$$

F-3.2.2. The coefficients a can be found for any sample size between 3 and 50 in Table B-19 of Appendix B. The value *k* is the greatest integer less than or equal to $n/2$.

F-3.2.2.1. Reject normality if the calculated statistic $W < W_\alpha$, where the critical values W_α are listed in Table B-20 of Appendix B.

F-3.2.2.2. If $W \geq W_\alpha$, do not reject the assumption of normality. Typically, one assumes the data are approximately normal for further statistical analysis.

F-3.3. <u>Example of Shapiro-Wilk W Test</u>. Consider using the Shapiro-Wilk to test the subsurface soil background chromium results for normality. The results (in mg/kg) are as follows: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, and 3.84.

F-3.3.1. Hypothesis test for Shapiro-Wilk *W* test:

H_0: The data are normally distributed.

H_A: The data are not normally distributed.

F-3.3.2. Estimate the sample standard deviation, $s = \sqrt{0.5255} = 0.7249$.

F-3.3.3. Compute Shapiro-Wilk test statistic *W*, where $n = 8$, $k = 8/2 = 4$ and $b = 1.859$:

$$W = \left[\frac{b}{s\sqrt{n-1}} \right]^2 = \left[\frac{1.859}{0.7249\sqrt{8-1}} \right]^2 = 0.9395 .$$

Using an α level of 0.05 and $n = 8$, we find the critical value, W_α, from Table B-20 to be 0.818. As $W > 0.818$, there is insufficient evidence to reject the assumption of normality.

$x_{(i)}$	$x_{(n-i+1)}$	$x_{(i)} - x_{(n-i+1)}$	$a_{(n-i+1)}$	b_i
3.84	5.86	2.02	0.6052	1.2200
4.26	5.74	1.48	0.3164	0.4683
4.53	5.29	0.76	0.1743	0.1325
4.60	5.28	0.68	0.0561	0.0381
5.28	4.60	–0.68		
5.29	4.53	–0.76		
5.74	4.26	–1.48		
5.86	3.84	–2.02		

F-4. Coefficient of Variation. The coefficient of variation (CV) may be used to quickly determine whether or not data follow a normal curve by comparing the sample CV to 1. However, the CV evaluation is not reliable. The use of the CV is valid only for some environmental applications if the data represent a non-negative characteristic, such as contaminant concentrations. If the CV is much greater than 1, the data should not be modeled with a normal curve. However, this method should not be used to conclude the opposite; do not conclude that the data can be modeled with a normal curve if the CV is less than 1. Furthermore, the sample CV $(s/\bar{x})$ can be greater than 1 when the population CV (σ/μ) is between 0.5 and 1. This is because of the sample CV being a random variable and estimating the true CV with some degree of error (EPA 68-W0-0025). 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. Details for estimating the CV are presented in Appendix D.

F-5. Range Tests.

F-5.1. General. Range tests for normality have been developed based on the knowledge that virtually 100% of the area of a normal curve lies within plus and minus 5 standard deviations from the mean. 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. These range tests are not as reliable as the previously discussed tests, and are recommended only if computer procedures or look-up tables for the other tests are not available. However, both range tests are relatively simple to use, so they are presented here.

F-5.1.1. The Studentized range test compares the range of the sample to the sample standard deviation. Tables of critical values for sample sizes up to 1000 (Table B-21 of Appendix B) are available for determining whether the absolute value of this ratio is significantly large.

F-5.1.2. Directions to conduct the Studentized range test and an example of this test follow in Paragraph F-5.2.

F-5.1.3. 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 easier to apply.

F-5.1.4. Alternatively, 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.

F-5.1.5. Directions for calculating Geary's test are presented below in Paragraph F-5.3

F-5.1.6. This test does not perform as well as the Shapiro-Wilk test or the Studentized range test. However, because Geary's test statistic is based on the normal distribution, critical values for all possible sample sizes are available. An example application of Geary's test follows in Paragraph 5-4.

F-5.2. Directions and an Example of Studentized Range Test.

F-5.2.1. Directions.

a. Calculate sample range (R) and sample standard deviation (s).

b. Calculate the ratio R/s.

c. Compare to the critical values for R/s given in Table B-21 (labeled *a* and *b*).

If the calculated value of R/s falls outside the two critical values, then the data do not follow a normal curve.

F-5.2.2. Example. Consider using the Studentized range test to determine if the subsurface soil background chromium results can be modeled using a normal curve. The results are (in mg/kg) as follows: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, and 3.84.

Sample range $R = 5.86 - 3.84 = 2.02$

Sample standard deviation $s = \sqrt{0.5255} = 0.7249$.

$$R / s = 2.02 / 0.7249 = 2.787 \ .$$

The critical values for *R/s* in Table B-21 for *n* = 8 and α = 0.05 are 2.50 and 3.399. As 2.787 falls between these values, the assumption of normality is not rejected.

F-5.3. Directions for Calculating Geary's Test. Calculate the sample mean $\bar{x}$, the sample sum of squares (SSS), and the sum of absolute deviations (SAD):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{SSS} = (n-1)\,s^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

$$\text{SAD} = \sum_{i-1}^{n} \left| x_i - \bar{x} \right| \ .$$

F-5.3.1. Calculate Geary's test statistic

$$a' = \frac{\text{SAD}}{\sqrt{n(\text{SSS})}} \ .$$

F-5.3.2 Test *a* for significance by computing

$$z = \frac{a' - 0.7979}{0.2123 / \sqrt{n}} \ .$$

Here, 0.7979 and 0.2123 are constants used to achieve normality.

F-5.3.3. Use Table B-15 of Appendix B 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. The statistic a' is sufficiently small or large to conclude the data are not normally distributed if $|z| > Z_{1-\alpha}$.

F-5.4. Example of Geary's Test. Consider using Geary's test to see if the subsurface soil background chromium results can be modeled using a normal curve. The results are (in mg/kg) as follows: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, and 3.84.

F-5.4.1. Calculate the sample mean $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}(4.60 + 5.29 + 4.26 + 5.28 + 4.53 + 5.74 + 5.86 + 3.84) = 4.925 \ .$$

F-5.4.2. Calculate the SSS:

$$\text{SSS} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

$$\sum_{i=1}^{n} x_i^2 = (4.60)^2 + (5.29)^2 + (4.26)^2 + (5.28)^2 + (4.53)^2 + (5.74)^2 + (5.86)^2 + (3.84)^2 = 197.73$$

$$\frac{(\sum_{i=1}^{n} x_i)^2}{n} = \frac{(39.4)^2}{8} = 194.05 \ .$$

So, $\text{SSS} = 197.73 - 194.05 = 3.68$.

F-5.4.3. Calculate the sum of absolute deviations (SAD):

$$\begin{aligned} \text{SAD} = \sum_{i=1}^{n} |x_i - \bar{x}| = |4.60 - 4.925| + |5.29 - 4.925| + |4.26 - 4.925| \\ + |5.28 - 4.925| + |4.53 - 4.925| + |5.74 + 4.925| \\ + |5.86 - 4.925| + |3.84 - 4.925| = 4.94 \ . \end{aligned}$$

F-5.4.4. Calculate Geary's test statistic:

$$a' = \frac{\text{SAD}}{\sqrt{n(\text{SSS})}} = \frac{4.94}{\sqrt{8(3.68)}} = 0.910 \ .$$

F-5.4.5. Test a' for significance by computing:

$$z = \frac{a' - 0.7979}{0.2123/\sqrt{n}} = \frac{0.910 - 0.7979}{0.2123/\sqrt{8}} = 1.49 \ .$$

Here, 0.7979 and 0.2123 are constants used to achieve normality.

F-5.4.6. Using Table B-15 of Appendix B to find the critical value $Z_{1-\alpha}$, where α = 0.05, then $Z_{1-\alpha}$ = 1.645. Since $1.49 \not> 1.645$, there is not enough information to conclude that the data do not follow a normal distribution.

F-6. Goodness-of-Fit Tests. Goodness-of-fit tests are not practical to do manually. Because these are included in most statistical software packages, detailed instructions for doing them are not included. Following is a brief overview of these tests with recommendations for their use.

F-6.1. Goodness-of-fit tests are used to determine whether data conform to some theoretical probability distribution. However, unlike the tests previously discussed, these tests can be used to see if a data set fits any specified probability distribution, not just the normal distribution. In contrast, the Shapiro-Wilk test can be used only to determine whether a data set is normally distributed.

F-6.2. There are many different goodness-of-fit tests. One classic test is the chi-square test, which partitions the data into groups, comparing these to the expected groups from a known distribution. There are no fixed methods for selecting these groups, and this test requires a large sample size because at least five observations per group are required to implement it. In addition, the chi-square test does not have the power of the Shapiro-Wilk test or some of the other tests mentioned. For these reasons, the chi-square test is not recommended.

F-6.3. Another way of using a goodness-of-fit test is based on the empirical distribution function. Empirical distribution functions estimate the true cumulative distribution functions underlying a set of data. An empirical distribution is generated from the data set and compared to the theoretical cumulative distribution. If the empirical distribution function is not close to the given cumulative distribution function, then there is evidence that the data do not come from that function.

F-6.4. Various methods have been used to measure the discrepancy between the sample empirical distribution function and the theoretical cumulative distribution function. These measures are referred to as empirical distribution function statistics. The best known of these is the Kolmogorov-Smirnov (K-S) statistic. The K-S approach is appropriate if the sample size exceeds 50 and if $F(x)$ represents a specific distribution with known parameters (e.g., a normal distribution with $\mu = 100$ and $\sigma^2 = 30$). A modification to the test, called the Lilliefors K-S test, is appropriate when $n > 50$ for testing that the data are normally distributed and when the $F(x)$ is based on an estimated mean and variance.

F-6.5. Unlike the K-S type statistics, most empirical distribution function statistics are based on integrated or average values between the empirical and cumulative distribution functions. The two most powerful are the Cramer-von Mises and Anderson-Darling statistics. Extensive simulations show that the Anderson-Darling empirical distribution function statistic is as effective as any, including the Shapiro-Wilk statistic, when testing for normality. However, the Shapiro-Wilk test is applicable only to a normal distribution, while the Anderson-Darling method is more general. Because it is unlikely that the user of this manual will ever need to use these tests, they will not be described further. When using a computer software package, a *p* value is typically given. If the *p* value is low (i.e., typically less than 0.01 to 0.1), then the assumption of normality is rejected.

Section II
Data Transformations

F-7. Introduction. Any mathematical function $f(x)$ that is applied to every point in a data set, x, is called a transformation (e.g., $Ln(x)$ is calculated for every data value x). For the transformation

$$y = f(x)$$

the values of x are the original data values and the corresponding values $y = f(x)$ are the transformed data values. An inverse transformation is a function, $f^{-1}(x)$, which, when applied to all of the transformed data values, results in the original data values:

$$f^{-1}(y) = f^{-1}[f(x)] = x \ .$$

F-7.1. For example, if $y = Ln(x)$, then $f^{-1}(y) = exp(y)$ because $exp[Ln(x)] = x$.

F-7.2. Data transformations are frequently done to obtain normally distributed data sets. By transforming the data, assumptions that are not satisfied in the original data can be satisfied by the transformed data. For example, a right-skewed distribution can often be transformed to be approximately Gaussian (normal) by using a logarithmic transformation or square root transformation. After a data set is transformed, graphical methods and statistical tests verify that the transformed data set is normal. If a transformed data set is normal, then statistical tests that rely on normality are performed using the transformed data. However, finding a transformation that results in a normal data set may be difficult. The selection of a suitable transformation will be dependent upon the nature of the data set and is beyond the scope of this document. Some commonly used transformations will be discussed but only lognormal transformation will be discussed in any detail.

F-7.3. A potential disadvantage of any transformation arises when it is necessary to interpret the results of the statistical evaluation in terms of the untransformed data. For example, in general, if the mean of the transformed data set is calculated, then this quantity will not correspond to the mean of the untransformed data set when an inverse transformation is performed. For example, as previously stated, if $Y = Ln(X)$ is normally distributed with a population mean (μ_Y) and population variance, σ_Y^2, then the mean (μ_Y) corresponds to the population median of X rather than to the population mean of X, μ_X. (Because $X_{p\,=\,0.5}$ is the mean of X and $Z_{0.5} = 0$ in Equation E-12, the median of X is equal to exp(μ).)

F-7.4. If a transformation is performed, inverse transformations to the original data set should be avoided. Decisions should be based upon the statistical analyses of only the transformed data. For example, assume that two different data sets are approximately normally distributed with similar variance after transformation. The objective is to determine whether the data sets are significantly different from one another (even though both data sets possess similar variances). The mean of the first transformed data set would be statistically compared to the mean of the second transformed data set. It would be inappropriate to perform inverse transformation for the two means (to express them in the original measurement units) prior to performing the comparison.

F-7.5. While transformations are useful for dealing with data that do not satisfy statistical assumptions, they can also be used for other purposes. 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. Transformations may also make the analysis of data easier by changing the scale into one that is more familiar or easier to
analyze.

F-8. Logarithmic. A logarithmic transformation may be useful when the original measurement data follow a lognormal distribution. Data may be lognormally distributed when the variance is proportional to the square of the mean (refer to Equation E-11) or, equivalently, when the coefficient of variation (ratio of standard deviation to mean) is constant over all possible data values:

$$\mathrm{CV} = \sigma_X / \mu_X = \text{constant.}$$

F-8.1. 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.

F-8.2. The logarithmic base (either natural or base 10) needs to be consistent throughout the analysis. However, it does not matter whether a natural (Ln) or base 10 (Log) transformation is used because the two transformations are related by a constant:

$$Ln(X) = 2.303\ Log(X).$$

F-8.3. The $Log(x)$ or $Ln(x)$ cannot be transformed when $x = 0$. This is usually not a problem for environmental applications because non-detects are not typically reported as zero but to some positive reporting (censoring) limit. If some of the original values are zero, it is customary to add a small quantity (ε) to make the data value non-zero, as the logarithm of zero does not exist. However, this introduces some error for the statistical evaluation. The size of ε depends on the magnitude of the non-zero data. It is recommended that the statistical evaluation be performed using several values of ε to determine if it is sensitive to the choice of ε. An initial value of one-tenth of the smallest non-zero value is recommended.

F-9. Square Root.

F-9.1. An overview rather than a detailed discussion of the square root transformation is presented here. The square root transformation may be used when the data values are 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. According to EPA's SW-846 methodology, if the mean and variance of a data set are equal, indicating data from a Poisson distribution, then the data can be transformed using a square root transformation so the data can achieve normality.

F-9.2. 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. The square root transformation may also be useful when developing control charts for intrawell comparisons when the assumption of normality is a concern. For further discussion on control charts, see Appendix Q.

F-10. Inverse Sine (Arcsine). An overview rather than a detailed discussion of the inverse sine transformation is presented here. 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). According to EPA's SW-846 methodology, if the mean is less than the variance of a data set, indicating data from a negative binomial distribution, then data can be transformed using an arcsine transformation to achieve normality. Special tables must be used to transform the proportions into degrees.

F-11. Box-Cox Transformations. An overview rather than a detailed discussion of the Box-Cox transformation is presented here. The Box-Cox transformation is a complex but useful transformation that takes the original data and raises each data observation to the power γ. Box-Cox is typically used in regression modeling (a statistical methodology used to identify the best-fitting equation for a set of data) and would be done using statistical software. Box-Cox is also performed when a data set is not normal, but it is desirable to produce normally distributed transformed data. A logarithmic transformation is a special case of the Box-Cox transformation. The Box-Cox family of transformations is defined as follows:

$$X^{\gamma} = \begin{cases} X^{\gamma}, & \gamma \neq 0 \\ Log(X), & \gamma = 0 \end{cases}$$

where γ is a parameter that defines the transformation (Hahn and Meeker, 1991).

F-11.1. Note both the logarithmic transformation and the square root transformation are simply Box-Cox transformations with $\gamma = 0$ and $\gamma = 0.5$, respectively. The parameter γ is generally unknown. The objective is to find a value of γ such that the transformed data are normally distributed and the variance is as constant as possible over all possible concentration values. In general, transformations with $\gamma < 1$ are applied to normalize positively skewed data, and transformations with $\gamma > 1$ are used to normalize negatively skewed data. The value of γ required to normalize the data decreases (from 1) as the degree of positive skew increases. For example, a transformation with $\gamma = 0.5$ might be applied for a distribution with a slight positive skew, and a value of $\gamma = 0$ (a log transform) might be applied for a more positively skewed distribution. From Hahn and Meeker (1991): “One may try different values of γ (i.e., $\gamma = 1, 0.5, 0.33, 0$, and -1, corresponding to no transformation, square root, cube root, log, and reciprocal transformations, respectively) to try to find a value (or range of values) that gives a probability plot that is nearly linear. In some cases physical considerations or experience may suggest such a value.”

F-11.2. Analytical methods are also available, such as the maximum likelihood technique, to find the optimal γ. A statistical software package would be used to find the value of γ for the best transformation; that is, the value of γ that produces the most normal data set once the transformation is applied. For example, if γ is nearly equal to zero (e.g., $\gamma = 0.03$), then a logarithmic transformation ($\gamma = 0$) would typically be selected and would produce a data set that is the most normally distributed relative to other Box-Cox transformations (such as a square root transformation). Statistical tests that require normality would subsequently be performed using the transformed data. However, as is true of any transformation, one of the disadvantages of Box-Cox is the difficulty in interpreting the transformed data in terms of the original measurement units.

Section III
Recommendations

F-12. General. Analysts can perform tests for normality with samples as small as three; however, the tests lack statistical power owing to the small sample size. For small sample sizes, it is recommended that a normal distribution not be assumed for the data and that a nonparametric statistical test, one that does not assume a distributional form of the data, be selected instead. Ideally, an adequate sample size to provide the necessary power for statistical tests will have been selected prior to data collection.

F-12.1. This document recommends using the Shapiro-Wilk *W* test wherever practical, along with a normal quantile plot and box-plot. The Shapiro-Wilk *W* test is one of most powerful tests for normality, and it is recommended in several EPA guidance documents as the preferred test when the sample size is less than 50. The Anderson-Darling statistic is also recommended (e.g., when available via statistical software). A normal quantile plot is helpful, no matter the sample size, to verify results from any test of normality. In practice, with the use of computers it may be possible to perform more than one fitness test, and determine which fit has the highest *p* value.

F-12.2. In general, with large sample sizes, both D'Agostino's test and the Shapiro-Wilk test will be overly sensitive to small deviations from lognormality or normality and will result in an unknown distribution assignment more often than is appropriate. In these cases, close examination of probability plots and the application of professional judgment in determining the appropriate distributional assumptions will be particularly important.

F-12.3. If the Shapiro-Wilk *W* test is not feasible, then using either Filliben's statistic or the Studentized range test is reasonable. Filliben's statistic performs similarly to the Shapiro-Wilk test. The Studentized range is a simple test to use; however, it is not applicable for nonsymmetrical 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. If critical values for these tests (for the specific sample size) are not available, then implementing either Geary's test or the Lilliefors Kolmogorov-Smirnoff test is reasonable. Geary's test is easy to apply and uses standard normal tables similar to Table B-15 of Appendix B, and is 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.

F-12.4. Statistical professional judgment based on normal probability plots and results of the statistical tests should be considered when identifying a data value's distribution. If the statistician's professional judgment suggests a different distributional assumption than that determined by the statistical test or tests, the alternative distribution may be assumed as long as the statistician provides a defensible rationale for this decision.

F-12.5. It should be stressed the Shapiro-Wilk *W* test is a good test to use to evaluate whether a set of data has been drawn from a normal or lognormal distribution. However, this test will not have very much power to reject the null hypothesis of normality or lognormality if the sample size is small.

F-12.6. In conclusion, results from tests regarding the assumption of normality should always be reviewed graphically.

F-13. <u>Data Fitting Multiple Distributions</u>. When data are found to fit more than one distribution, there are a few things to consider in making a decision about which distribution would be most appropriate. One thing to consider is the *p* value. After running a test of distributional assumptions (Shapiro-Wilk, chi-square, Kolmogorov-Smirnoff, etc.), it would be appropriate to use the distribution that had the higher *p* value. Consideration should be given to the sample size of the data; data containing just a few samples may not provide enough information about the true distribution.

F-13.1. Another thing to question is the purpose of identifying the data's distribution. If it is to verify a distributional assumption for a statistical test and the data fit multiple distributions, it may be appropriate to perform the test using several statistical methods and evaluate results from each to see what can be learned. If a distributional assumption is needed to estimate a confidence interval or upper confidence limit, then it may be appropriate to identify which distribution would provide the more conservative estimate.

F-13.2. It is often difficult to interpret the results of statistical tests conducted on transformed data in terms of the original units to make these types of comparisons. If transformation produces only a slightly larger *p* value, it seems advisable not to perform the transformation. For example, if data follow a normal and lognormal distribution, a lognormal UCL can be quite larger than the normal UCL estimate owing to the inherent nature of a lognormal distribution. If the UCL should be used to evaluate risk at a site, a lognormal UCL would provide the more conservative estimate of risk.

APPENDIX G

Detection Limits and Quantitation Limits

G-1. Introduction.

G-1.1. Environmental statistical analysis is complicated by a practical constraint on laboratory analysis—the technical impossibility of identifying zero concentrations. This means that it is physically impossible for a laboratory analysis to confirm the complete absence of the chemical or compound of interest. A chemical may be present at some unknown concentration below the low end of the concentration range that the analysis is able to report detect. Therefore, for most statistical applications that evaluate site data, there is a need to substitute some number (a "censored" value) that represents the lowest concentration reasonably detected. This threshold or censoring limit is often termed a "detection," "quantitation," or "reporting" limit. However, this Appendix provides separate definitions for the terms "
detection" and "quantitation limit" and does not use these terms interchangeably.

G-1.2. To determine which censoring limit should be used for statistical evaluations, it is necessary to understand how environmental laboratories define detection and quantitation limits, as these quantities are used to establish censoring limits. Unfortunately, the subject of detection and quantitation limits is often confused by the highly diverse, and often overlapping, definitions applied to these quantities. Furthermore, no standard approach to establishing censoring limits for environmental data exists. This Appendix describes some of the methods for establishing detection limits and subsequent requirements for substituting values for non-detects in the data set.

G-2. Detection Limits. No instrumental method of chemical analysis is capable of "seeing" a value of zero. All measurement systems are subject to bias and variability. A fundamental contributor to this is the presence of "noise" in the measurement process. Noise can have any number of sources. For example, if one examines the pictorial output from a gas chromatographic analysis (a chromatogram) of a control sample at the normal scale at which it is displayed in a commercial data package, one would observe a Gaussian peak that represents the analyte of interest and what appears to be a straight, smooth line beyond the peak referred to as the "baseline." Figure G-1 depicts a cartoon example. However, that same graph examined at a higher level of magnification would reveal a very different picture of fluctuations across the same line (Figure G-1). Those fluctuations constitute noise and can result from such factors as vibration in the environment around the instrument, fluctuations in electrical current or voltage, the incidental presence of contaminants in the system, or even stray ionizing radiation from universal background.

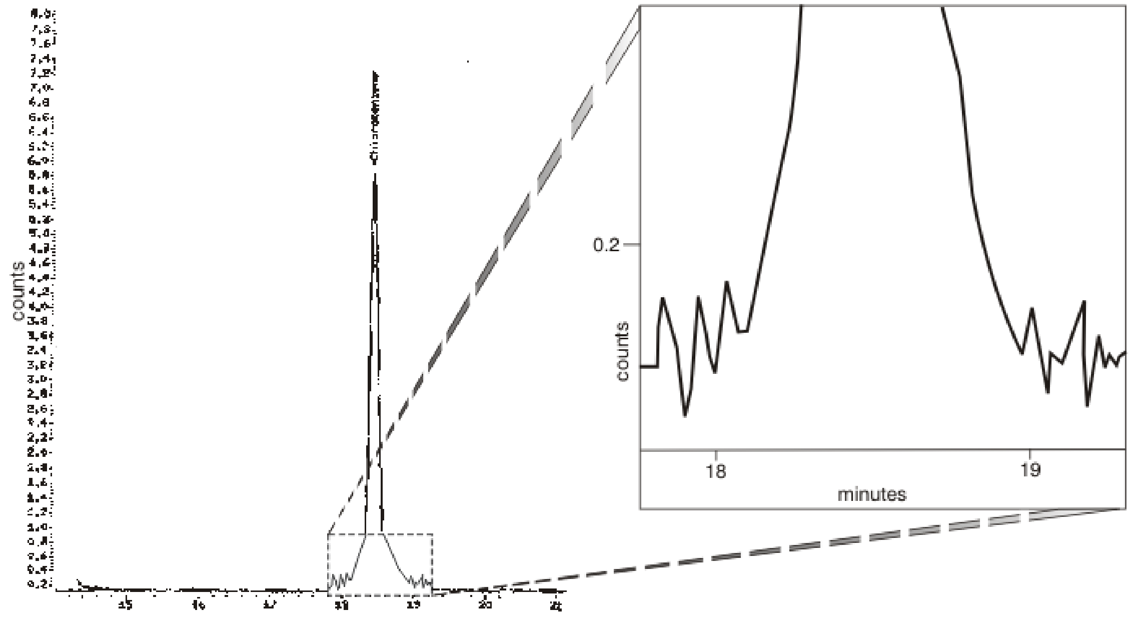


Figure G-1. Noise in GC Baseline.

G-2.1. If a very small amount of a target analyte were placed in the measurement system, assuming that the instrument was functioning properly, the analyte would cause a response in the detector that would be translated into a small Gaussian type peak on the chromatogram. However, as the concentration is decreased, the size of the peak decreases until it is "lost" in the noise of the measurement system. Because the amount of noise in the system at any given moment is essentially random, the amount of analyte that can be hidden by the noise is variable but, on average, is always greater than zero.

G-2.2. As the term is typically used in the environmental testing industry, a "detection limit" (DL) is the concentration that gives rise to an analyte peak or signal that is statistically greater than the surrounding baseline noise at a high level of confidence (typically the 99% level of confidence). The analyte cannot be confidently reported as present when the analyte concentration is less than the DL. Concentrations greater than the DL are reported as "detected."

G-2.3. However, theoretically, there are two types of "detection limits": The "Type I DL" that minimizes false positives (Type I error) and the "Type II DL" that minimizes false negatives (Type II error). A false positive occurs when an analyte is absent, or the true concentration is less than the baseline noise but is erroneously reported as present. A false negative occurs when an analyte is erroneously reported as less than or equal to some concentration when it is actually present at a greater concentration. The two types of detection limits are illustrated in Figure G-2.

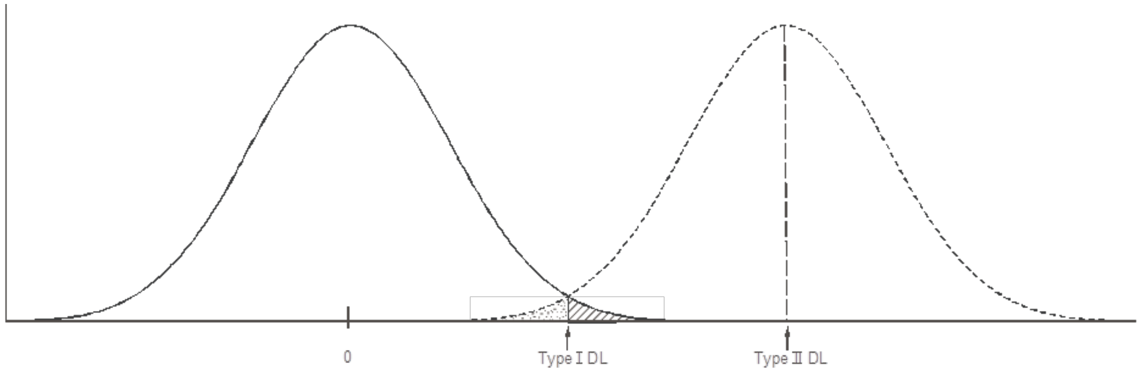


Figure G-2. "Type I DL" (L_C) and "Type II DL" (L_D).

G-2.4. The International Union of Pure and Applied Chemistry (IUPAC), an international, non-governmental organization that supports the advancement of chemical science, refers to the "Type I DL" as the "critical value" and the "Type II DL" simply as the "detection limit." Therefore, for simplicity and to conform with international nomenclature, the IUPAC terminology is predominately used in this document. The critical value is the threshold of analyte or instrument signal attributable to the presence of analyte that is statistically different from zero or baseline noise at a high level of confidence. The 99% level of confidence is used for chemical analyses. When an analyte is reported at a concentration greater than the critical value the conclusion is as follows: The analyte is present at some concentration greater than zero at the 99% level of confidence. The "detection" of the analyte is reported. However, if the analyte concentration reported from a measurement is less than the critical value, the analyte may or may not be present (the true analyte concentration may or may not be greater than zero). Under these circumstances, no conclusion regarding the presence or absence of the analyte is possible. The IUPAC detection limit is established to addresses "non-detections" of the analyte.

G-2.5. When a measurement is taken and the analyte is less than the critical value, the conclusion is that the analyte, if present, is present at some concentration less than the detection limit; the non-detection is reported as "less than the detection limit."

G-2.6. Currie's (1968) approach readily illustrates the nature of the critical value and detection limit on a conceptual level. Currie defines the critical level, L_C, as the concentration at which the binary decision of detection can be made with a specified level of confidence. The shaded area to the right of L_C in Figure G-2 represents the Type I error (i.e., the probability of concluding the analyte is present when the true concentration is zero). Currie defines the limit of detection, $L_{D,}$ to provide an acceptable Type II error rate. The shaded area to the left of L_C represents the Type II error (e.g., the probability of failing to detect the analyte when the true concentration is L_D). In order to calculate quantities L_C and L_D, the following simplifying assumptions are made: The concentrations are normality distributed, the standard deviation is known (or there is negligible uncertainty for the standard deviation), and the standard deviation is not a function of concentration and the “true” (population mean) concentration is zero. For the 99% level of confidence:

$$L_C = 2.33\sigma$$

$$L_D = L_C + 2.33\sigma = 2\ L_C$$

G-2.7. Unfortunately, it is common practice for environmental chemists to refer to the critical value as the “detection limit.” For example, the method detection limit (MDL), defined by 40 Code of Federal Regulations (CFR) Part 136 (Appendix B), is essentially a critical value (as defined by the IUPAC). There is no standard terminology for the IUPAC detection limit for environmental testing. There is a host of terminologies applied to detection and reporting limits depending on the source and the details of the definition. (QSM). Version 5 of the QSM measures analytical sensitivity in terms of the “Detection Limit” (“DL”), “Limit of Detection” (“LOD”) and “Limit of Quantitation” (“LOQ”). Conceptually, the “DL” is a “Type I DL” and the “LOD” is a “Type II DL.” The “LOD” is established and verified at least quarterly (for each environmental analyte and matrix) by processing laboratory control samples spiked at that concentration. The “LOQ” is discussed in G-4.

G-3. EPA Method Detection Limit and Other Detection Limits. There are two major DL estimators: those based on a “single concentration design” and “calibration designs.” The major disadvantage of single concentration designs is they assume that variability at a given concentration is constant (i.e., the variability near the DL is similar to that at higher concentrations). Typically, for a single concentration design, a set of replicate samples containing the analyte of interest at a fixed, known concentration are processed to calculate the critical value. Therefore, the critical value is determined at the single concentration for the replicate study and it is assumed that a higher or lower concentration would produce substantively the same value. The MDL is based upon a single concentration design. In calibration designs, the critical value is calculated using multiple concentrations over the range of the critical value. The multiple concentration levels provide a means to model the variance (e.g., or standard deviation) as a function of concentration. In this way, the resulting critical value estimate is not simply a function of sample spike concentration. However, single concentration designs are advantageous relative to multi-concentration designs because they are much simpler and less costly to perform. The critical value can be defined in many different ways; however, only the most commonly accepted method, the EPA MDL procedure, is discussed in detail.

G-3.1. EPA Method (Single Concentration Design). Historically, EPA has used single concentration designs, even though single concentration designs and their associated DL estimators are rarely completely justified. The MDL (defined by 40 CFR) is a single concentration design for the critical value that most environmental testing laboratories use.

G-3.1.1. The EPA defines an “instrument detection limit” (IDL) as an experimentally derived quantity arrived at by repeatedly injecting a small but visible amount of a pure analytical standard into the instrument, measuring the variability in the quantitative results, and calculating the IDL assuming 99% confidence that the observed response is not a false positive. The IDL is generally only performed for inorganic metals analyses. The IDL is typically calculated in the same manner as the MDL, using a Student's *t*-statistic. The two quantities differ predominately in the way the samples are processed. The IDL is determined via the direct instrumental analysis of standards containing the analyte of interest. However, when environmental samples are analyzed, they generally are not directly injected into instruments but are subject to a variety of prior preparatory processes (such as extractions, derivatizaton, solvent exchanges, cleanup, and dilutions). Each step in the processing adds additional noise or uncertainty to the measurement system, which the IDL calculation does not take into account. Therefore, IDLs tend to be smaller in concentration than the corresponding MDLs when samples are subjected to an extensive preparatory process prior to analysis. The minimum quantity of practical importance in environmental analysis is that amount that can be reliably distinguished from the sum of all the various sources of noise involved in the analytical method, the method detection limit (MDL). Thus, environmental laboratories typically use the MDL to characterize detection capability.

G-3.1.2. Although the MDL (as defined in 40 CFR) strictly applies to water matrices, it is applied to a broad range of analytical methods, including those for solid samples. This single concentration design requires a complete, specific, and well-defined analytical method. It is essential for all sample-processing steps of the analytical method to be included in the determination of the method detection limit. MDLs depend upon the sample preparatory procedures and the specific laboratory instrument used.

G-3.1.3. The EPA procedure used to estimate the detection limit is summarized below.

G-3.1.3.1. Prepare a homogeneous matrix that is free of analyte (e.g., reagent water or clean sand).

G-3.1.3.2. Prepare each sample mixture at a concentration of at least equal to or in the same concentration range as the estimated MDL in the matrix of interest.

G-3.1.3.3. Prepare a minimum of seven aliquots of the sample to be used to calculate the MDL and process each replicate through the entire extraction/digestion and analytical method.

G-3.1.3.4. Calculate the variance (s^2) and standard deviation (s) of the replicate measurements.

G-3.1.3.5. Calculate the MDL, using the formula: MDL = $t_{0.99.\nu}\, s$, where $t_{1\text{-}\alpha.\nu}$ is the Student's t value appropriate for the 99% confidence level with $\nu = n - 1$ "degrees of freedom"; and the number of measurements, $n \geq 7$. (The appropriate value of Student's *t* is typically found in a statistical table, and is equal to about 3.14 for *n* = 7 for the 99% level of confidence).

G-3.1.3.6. Review results to verify the reasonableness of the calculated DL.

G-3.1.4. The use of the MDL for decision-making (e.g., determining environmental impacts) has recently triggered intense scrutiny of the viability of the MDL for measuring detection capability. The following is a partial list of potentially flawed assumptions or problems associated with the MDL as defined in 40 CFR.

G-3.1.4.1. The MDL addresses false positives (i.e., Type I error), but does not address false negatives (Type II error); for example, a non-detection cannot be confidently reported as "< MDL." (However, it should be noted that there is controversy regarding the interpretation of the MDL in terms of the IUPAC definitions of the critical value and detection limit; some individuals have argued that the MDL is actually an IUPAC detection limit.)

G-3.1.4.2. The MDL underestimates method variability as it is typically calculated using a small number of replicates within a short period of time and has been interpreted to be a prediction limit for the next single future observation, minimizing false positives at the 99% level of confidence for only one future environmental sample (and not a set of multiple samples) when the analyte is absent (though it should be noted that the interpretation of the MDL as a prediction limit is also controversial).

G-3.1.4.3. The standard deviation is assumed to be constant (i.e., not a function of concentration).

G-3.1.4.4. Normality is assumed.

G-3.1.4.5. No analytical bias is implicitly assumed (e.g., no analyte loss, average analyte "recoveries" of 100%). (The MDL accounts for analytical method variation in the form of random "precision error.")

G-3.1.4.6. The matrix used to perform the MDL study (e.g., reagent water) is assumed to be equivalent (with respect to all physical or chemical properties that would affect detection capability) to the actual environmental matrices that will be tested (e.g., waste water and groundwater).

G-3.1.5. In general, one or more of the assumptions discussed above are routinely violated to some extent for environmental testing. MDLs are statistically derived quantities and are only estimates of the actual detection limit (critical value). For example, based on purely statistical considerations, MDLs are uncertain by a factor of approximately two. Furthermore, because MDLs are typically generated by processing clean material (such as purified water or sand) rather than actual environmental samples, they represent "best case" detection capability. In general, the material analyzed to calculate the MDLs is not representative of the chemical and physical composition of the environmental samples. Detection limits calculated using an actual environmental matrix could be higher than the MDL by an order of magnitude. However, because of these factors, environmental laboratories often report “detection limits” several times greater than MDLs (although there is no uniform standard for how this is done). The detection limits proposed in Paragraph G-3 overcome the first two shortcomings of the MDL discussed above.

G-3.1.6. Lastly, when detection limits such as the MDL are constructed from prediction limits (using either a single concentration or calibration design), in order to minimize false positives at the specified level of confidence, a new detection limit must (in theory) be calculated (from a new study) prior to each new sample being analyzed. However, this is not done in practice. Detection decisions for an enormous number of test samples are calculated based on the results obtained from a single MDL study. This results in a much greater frequency of false positives than 1%. To ensure that false positives are minimized for a large unspecified number of future measurements, detection limits may be constructed from tolerance intervals so that a large proportion of future measurements, *p*, will be less than the upper tolerance limit (UTL) with a high level of confidence when the “true” concentration is zero. For the critical value, an UTL for $p100\%$ coverage (e.g., where $p = 0.99$) at the $(1 - \alpha)100\%$ (e.g., 99%) level of confidence could be constructed for a “true” concentration of zero (e.g., refer to Paragraph G-3).

G-3.2. Alternative Method (Single Concentration Design). Although the Currie approach is conceptually viable, there is a major practical problem with the approach. Currie did not propose a practical experimental design to calculate L_C, but expressed L_C in terms of the population standard σ (which is usually unknown), rather than the sample standard deviation, *s*. (In other words, $L_C = 2.33\sigma$ only when the distribution is normal and σ is known.) Similarly, L_D cannot be calculated using σ if this quantity were unknown. However, for a normal distribution, L_C can be defined as an upper tolerance limit for a population mean $\mu = 0$ and can be calculated from *s* using an equation of the form (Georgian and Osborn, 2003):

$$L_C = K_{p,1-\alpha,n-1}\, s$$

The standard deviation *s* is calculated from a set of *n* replicate samples (e.g., a clean matrix such as reagent water spiked with the analyte of interest) that are processed through the entire analytical method. The factor $K_{p,1-\alpha,n-1}$, which depends upon the *coverage probability* (*p*), level of confidence $(1 - \alpha)$ and number of samples (*n*), can be calculated from Tables B-2 and B-15 using the following equation:

$$K_{p,1-\alpha,n-1} = Z_p \sqrt{[(n-1)/\chi^2_{n-1,\alpha}]} \ .$$

For example, if $1 - \alpha = 0.95$ (i.e., $\alpha = 0.05$), $p = 0.99$ and $n = 7$, then from Table B-2,

$$\chi^2_{n-1,\alpha} = \chi^2_{6,0.05} = 1.635$$

and, from Table B-15, $Z_p = Z_{0.99} = 2.33$. Therefore,

$$K_{0.99,0.95,6} = 2.33\sqrt{[(7-1)/1.635]} = 4.46 \quad .$$

If a large number of blank samples are analyzed, with 95% confidence, at least 99% of all the measurements will be less than $L_C = 4.46\ s$. The above equation, however, assumes normality and constant variance. A conservative approximation for L_D would consist of initially calculating L_C using the equation above then setting L_D equal to two times L_C.

G-3.3. Calibration Designs. In one type of calibration design, a series of samples are spiked at different known concentrations in the range of the hypothesized critical value, and variability is determined by examining the deviations of the actual response signals from a fitted regression line (instrument response versus concentration). In this design, it is typically assumed that the distribution of the deviations from the fitted regression line is normal with constant variance across the range of concentrations used for the study. The relationship between response signal (Y) and spiking concentration (X) in the region of the critical value is assumed to be a linear function of the form:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where the (population) "residual" $\varepsilon = Y - (\beta_0 + \beta_1 X)$ is the deviation of the measured value of Y from the "true" regression line $\beta_0 + \beta_1 X$. It is assumed that the distribution of values for ε is normal with mean $\mu = 0$ and some constant variance. A set of n measurements (x_i, y_i) would be used to estimate a line of the form $Y = b_0 + b_1 X$, where the sample slope, b_1, estimates the population parameter β_1 and the sample intercept, b_0, estimates the population parameter β_0. The regression model is used to calculate the critical value and detection limit by constructing either prediction or tolerance limits for the regression line, $Y = b_0 + b_1 X$. (The specific mathematical formulas used are beyond the scope of this document.)

G-3.3.1. Hubaux and Vos method calibration design is an example of an approach in which statistical prediction limits are used to calculate DLs. The critical, L_C, value is calculated from a 99% prediction interval for the linear regression model. A single future measurement will be less than L_C at the 99% level of confidence when the "true" concentration is zero. The limit of detection, L_D, is then defined as the smallest concentration at which there is 99% confidence a value greater than L_C will be obtained. This method assumes that the variability is constant throughout the range of concentrations used in the calibration design (e.g., if this assumption is violated, a variance stabilizing transformation might be applied and the assumption of constant variance may be reevaluated). The critical value obtained from the Hubaux and Vos design can be viewed as a multi-design concentration version of the single concentration-designed MDL (e.g., since the MDL is also a prediction limit, minimizing false positives for only one single future observation). Regression models used for multi-concentration designs can also be used to define detection limits based on prediction and tolerance intervals. A tolerance or prediction interval can be constructed for each possible value of the independent variable *X*.

G-3.3.2. As previously stated, the Hubaux and Vos calibration design assumes that the variance is homogeneous (constant) throughout the range of calibration function. This assumption is rarely completely justifiable. In practice, variation in the response signal is often proportional to the concentration. For example, if violations of this assumption are ignored, the variability at low levels can be overestimated and, as a result, detection limits can be overestimated. However, some calibration designs account for non-constant variance. For example, the detection limits for non-constant variance calibration designs can be calculated using a technique called weighted least squares (WLS). The WLS calibration design is similar to the Hubaux and Vos design, but the underlying regression model would assume, for example, that variance is proportional to concentration (Gibbons and Coleman, 2001).

G-4. Quantitation Limits. The ability to distinguish between the presence or absence of an individual analyte, particularly in a complex mixture such as an environmental sample, does not imply the ability to accurately and precisely measure the quantity of analyte present in the mixture. Imagine, for example, a peak partially hidden in the noise of an instrument. If the quantity of analyte is measured as proportional to the height or area of the response, as is the usual case in environmental analysis, from what point is it measured? Where is the baseline? Should it be measured from the lowest point in the noise, the average noise level, or the top of the noise? In other words, because the baseline is constantly shifting, what portion of the observed peak is noise and what portion is response? The magnitude of the response ascribable to the analyte (e.g., peak area) cannot be known with a high degree of certainty (high accuracy and precision); therefore, the measured value must, by definition, be equally suspect. There is a point at which the measured value is so much larger than any possible contribution from measurement noise that the noise becomes negligible relative to the analyte result. That point is the quantitation limit (QL). However, there is no standard terminology for this quantity in the environmental testing industry. It could be referred to as a “report limit” or erroneously referred to as “detection limit.” Terms such as “practical quantitation limit” or “contract required quantitation limit” could be used. Furthermore, as used by environmental testing laboratories, these terms may, but not would necessarily, refer to the “quantitation limit” as it is defined in this document.

G-4.1. In EPA terminology, the QL is, by definition, a value sufficiently removed from the detection limit to ensure that quantitative statements made at that value meet defined degrees of precision and accuracy by most laboratories under most analytical conditions. Because the definition is vague, the QL is also vague. In fact, most practical applications of this concept are altogether arbitrary. For example, in EPA SW-846, the EQL for a given analysis is defined as 5 to 10 times the MDL. However, the multiplication factor is somewhat arbitrary (e.g., various definitions of the QL for various programs have required the MDL to be multiplied by factors ranging from 2 to 10). Some justification for the use of a factor of 5 to 10 is as follows: If the MDL is assumed to be roughly equal to the magnitude of the uncertainty from analytical noise, the relative error should be 20 to 10% at 5 and 10 times the MDL, respectively. However, it should be noted that this assumes that analytical bias is negligible and the standard deviation (used to calculate the MDL) is not a function of concentration and possesses negligible uncertainty. In general, these are not valid assumptions. In particular, the standard deviation is typically an increasing function of concentration and can vary by a factor of about two when it is calculated at a fixed concentration using only seven replicates (as in 40 CFR). Setting the QL at a concentration at least 5 or 10 times the MDL is stated only as guidance (e.g., since the uncertainty at these levels may still be relatively large).

G-4.2. The “Practical Quantitation Limit” (PQL) is defined as the lowest limit of quantitation achievable by laboratories within specified limits on precision and accuracy during routine laboratory operating conditions. Unfortunately, acceptance limits for precision and accuracy at the PQL are seldom defined. In practice, the PQL is typically established by multiplying the MDL (as derived from 40 CFR Part 136 instructions) by a factor of three to five (from EPA SW-846, Chapter 1). The result obtained is the EQL. The EQL, being a multiple of the statistically derived MDL, will be different for each analyte tested. In the commercial laboratory community, PQLs are frequently set at the low point of the curve and are relatively uniform for methods where multiple analytes are simultaneously determined. The values thus obtained are variously referred to as PQLs, Reporting Limits (RLs), Less Than (< or LT), Non-Detects (NDs), or “U”- values.

G-4.3. To ensure acceptable precision and accuracy at any arbitrarily defined QL, quality control samples spiked at the QL could be included in the analytical sequence to actually measure the precision and accuracy of the measurement process (e.g., using control charts). Thus, this approach would quantify the uncertainty at the QL for “clean matrices” and is predominately how the DoD QSM defines the “Limit of Quantitation” (“LOQ”). Precision and bias at the “LOQ” are determined by processing laboratory quality control samples at that concentration. The “LOQ” is periodically verified by analyzing these low-level quality control samples at least quarterly. The QSM also requires the “LOQ” to fall within the calibration range of the analytical method, as instrumental response is typically unknown at concentrations less than the lowest initial calibration standard.

APPENDIX H

Censored Data

H-1. <u>Introduction</u>. Laboratories report analytical data in two ways, as censored or uncensored. An environmental testing laboratory reports a result as “non-detected” or “ND” when the result is below some numerical reporting threshold. The non-detect is typically reported as “< *X*” (e.g., or “*X* U”) where *X* is some numerical value. This is called a “censored result,” and the value of *X* is called the “censoring limit.” Results reported as “detected” are “uncensored” results. Typically, uncensored results are numerical values in concentration units that are greater than either the critical value or the censoring limit. Unfortunately, different environmental laboratories use different types of censoring limits and reporting conventions. There is no standard industry practice regarding how to establish the censoring limit for non-detections. To exacerbate matters, as discussed previously, there is no standard terminology for the censoring limit. Reporting conventions differ from laboratory to laboratory. Some laboratories refer to the censoring limit as the DL, while other laboratories refer to this value as the “reporting limit” (RL).

H-1.1. Before evaluating censored data, it is important to understand the nature of the censoring limit being used, that is, to understand how it is being defined for a particular set of data. To confidently report a non-detect at the censoring limit, the censoring limit must be equal to or greater than the detection limit (as this quantity is defined by the IUPAC); ideally, non-detects should be reported as “< DL” (larger values are undesirable for statistical evaluations and smaller values are undesirable for the minimization of minimize false negatives). For normally distributed data, in general, the censoring limit should be at least two times greater than the reported critical value. However, it is not uncommon for laboratories to report non-detects to values as low as the MDL (where false negatives cannot be reliably reported). The censoring limit is often the laboratory’s practical quantitation limit (PQL), which may also be simply called the QL. Under these circumstances, a laboratory reports numerical results greater than the QL as quantitatively reliable values. A result less than the QL may be reported as detection, consisting of a numerical value with a “data qualifier” or “flag” if the result is greater than the critical value (e.g., the method detection limit), or the result may be reported as a non-detect as “< QL.” For example, if QL = 10 ppm, MDL = 1 ppm, and a result of 5 ppm is measured, the laboratory may report the result as either “5 J” or as “< 10.” The reporting of the result as “5 J” indicates that the analyte is present, but the concentration of 5 ppm is a highly estimated value (i.e., is not quantitatively reliable). If the result were reported as “< 10,” the result would be a censored value (indicating the concentration of analyte is no greater than 10 ppm). The J-qualifier is typically applied when the analyte is believed to be present at some concentration less than the QL. (Detection and quantitation limits are discussed in detail in Appendix G.)

H-1.2. When measurement data are reported as “ND,” the exact concentration of the chemical is unknown, but lies somewhere between zero and the censoring limit. No quantitative information is available for a non-detect (except that the result is less than the censoring limit) because no estimate is provided to quantify how much smaller the result is than the censoring limit. Although useful for data reporting and presentation, censored data complicate statistical analyses and data interpretation. Qualitative results cannot be used because statistical calculations require numerical values rather than attributes. For example, the inequality "< 10 ppm" cannot be substituted into the equation to calculate the sample mean although a value of 5 could be substituted for a result reported as "5 J."

H-1.3. Statistical literature, Federal standards, and USEPA guidance advocate the use of uncensored measured concentrations for statistical calculations. Uncensored data give rise to more accurate estimates of mean and standard error than censored data, which result in more accurate data interpretation and more reliable conclusions. However, under these circumstances, numerical values (even negative values) would be reported for each sample regardless of the magnitude of the concentrations relative to the DLs. Unfortunately, in practice, censored data are typically reported for environmental applications because uncensored data are often unavailable or difficult to obtain, especially for prior sampling events (e.g., some laboratory instruments are incapable of reporting uncensored values). Requesting uncensored data may also increase analytical laboratory costs because uncensored data are not routinely reported, but it can be done at a reasonable cost for select analytical methods (e.g., typically, for metal analyses).

H-1.4. As censored data are commonly reported for environmental testing, the next Paragraph presents a variety of strategies for treating censored data. Some are recommended, while others should be used with greater caution. Gilbert (1987) and Gibbons (1994) contain more information on dealing with censored data. Helsel (2005) presents a number of useful statistical methods for censored environmental data that are strongly recommended. The statistical methods described in this Appendix are not as comprehensive or powerful as those described in Helsel (2005); they are presented to facilitate only a basic understanding of how to process censored data.

H-2. Overview of Strategies for Treating Censored Data. There are several possible approaches for treating censored data. Four general strategies are listed below and then described in more detail:

a. The censored values can be ignored (omitted from the statistical calculations).

b. Proxy values (e.g., the censoring limit, one-half the censoring limit, or zero) can be substituted for the NDs to obtain numerical values for computations (e.g., for the mean and variance).

c. Statistical quantities such as the mean and variance can be adjusted based upon the proportion of NDs by making certain distribution assumptions.

d. Nonparametric methods can be used.

No single approach can be used for all data sets and all data quality objectives. The characteristics of the data set and its end use must be taken into account when selecting the most appropriate approach.

H-2.1. Approach 1. The first approach, omitting the NDs from the data set, is typically undesirable as it decreases the total number of data points and the reliability (power) of the statistical evaluations. In addition, the NDs often provide valuable information about the environmental population of interest. For example, a set of NDs that are all less than some risk-based decision limit provides valuable information about the site. This approach is potentially viable only under select circumstances and for select data quality objectives. For example, it may be appropriate if there are a large number of samples for a study area and the censoring limit is small relative to some risk-based decision limit to which monitoring is being performed. If a statistical evaluation using only the set of detections were to indicate that contamination is present at concentrations significantly less than the decision limit, the omission of the NDs would probably not affect decision-making.

H-2.2. Approach 2.

H-2.2.1. The second approach is called the “substitution method.” Proxy or surrogate values are assigned to all the NDs. One approach for assigning proxy values is to assume that any value between zero and the censoring limit is equally probable and substitute one-half the censoring limit (midpoint of the range of possible values) for each ND. Other common proxy values are zero or the censoring limit itself. However, assigning proxies requires assumptions about the distribution of NDs. For example, assuming that all values less than the censoring limit are equally likely is equivalent to assuming a uniform probability distribution for all possible measurements between zero and the censoring limit. Assuming that all non-detects are equal to a fixed proxy value can bias the estimated standard deviation for the data set, particularly when a substantial number of results are NDs (see ASTM D-4210-89 for further discussion of this topic). For example, substituting the censoring limit could result in a sample mean that is biased high, and substituting zero could result in a mean that is biased low. Substituting one half the censoring limits may not bias the mean, but often adversely affects the estimate of the standard deviation. Biasing such summary statistics may result in erroneous conclusions about project objectives. In general, it is undesirable to assign proxy values, especially when a significant portion of the data set (e.g., more than 15%) contains censored values.

H-2.2.2. As noted previously, laboratories often report uncensored data below the censoring limit as estimated positive detections (commonly indicated as J-flagged values). Using these uncensored data for statistical computations (not necessarily for data reporting) prevents the need to assign proxy concentrations based on arbitrary algorithms

(EPA 9285.7-09A, Gilbert, 1987). While measurements below the censoring limit may not indicate the presence of target analytes as reliably as measurements above the limit, in many cases uncensored measurements are still better estimates of contaminant concentration than any proxy that might be applied. Generally, this approach allows data users and decision-makers to better characterize site conditions. Censored data are always relevant for determining the presence or absence of a contaminant at a site, as long as appropriate qualitative identification criteria have been satisfied.

H-2.3. Approach 3. The third approach entails adjusting the average and standard deviation instead of estimating proxy values for each ND result. However, to do this, it is also necessary to make assumptions about the data distributions (such as, all NDs vary in a manner similar to results above the censoring limit—maximum likelihood estimation procedure and the probability plotting method—or Cohen's method, which assumes a normal distribution). Adjustment methods provide accurate results only when the distribution assumptions are valid; otherwise, elevated estimates of the average and standard error could result. Usually, adjustment methods should be used when 15 to 50% of the values of the data results are censored.

H-2.4. Approach 4. A nonparametric approach should be considered when a significant portion of the data set consists of censored values. This approach typically involves ordering the data values (from smallest to largest) and replacing the data values with the corresponding rank number. The NDs are then treated as tied ranks and would be replaced by some common mid-rank value. Though not generally recommended, according to EPA guidance, if the DLs are not the same, then the NDs, instead of being treated as tied values, would be ranked according to their numerical estimates (EPA 68-W0-0025). The advantage of a non-parametric approach over the strategy of assigning proxy values is that no distribution assumptions are made. However, a larger number of data points are required for non-parametric methods to achieve the same level of confidence as parametric methods. Furthermore, though non-parametric methods can tolerate a greater proportion of NDs than parametric methods, non-parametric methods will not be viable if there are many NDs. For example, the median (refer to Appendix D) could not be determined from a data set that consists of more than 50% NDs.

H-2.5. Complicating Factors. For most projects, uniform numerical censoring limits will be available; however, there are instances when this is not the case. A laboratory can provide sample-specific detection limits or critical values (i.e., limit adjusted by the sample-specific dilution factor, soil moisture, or other analytical adjustments) that vary from sample to sample. In this case, use the sample-specific limits to establish the proxy values. As there is no standard nomenclature or well-established conventions for generating censoring limits in the environmental testing industry, it is recommended that the project chemist be consulted to establish the nature of the censoring limits being reported.

H-2.5.1. Censored results are sometimes reported as “ND” without the associated censoring limit. When censoring limits are not provided with data, this information can usually be obtained by contacting the laboratory if the analyses are current. If this information is not available, it might be viable to estimate a censoring limit based upon the lowest reported concentration, such as the lowest J-flagged result. Because J-flagged results are, by definition, concentrations that exceed the critical value, the minimum result represents a value that is closest to the critical value. A chemist should be consulted to examine the J-flagged values to determine if there are anomalous values that would set proxies at inappropriate levels. For example, an examination of the J-flagged results may indicate that there may be, in effect, two different censoring levels—one for “dirty” samples and one for “clean” samples. The project chemist might want to consider the issues of aliquot sizes and dilution conformity, among others factors, prior to making a final recommendation.

H-2.5.2. When using a nonparametric method to address NDs, ranking the data is often problematic when there are multiple censoring limits. For example, in general, it cannot be concluded that “< 10” represents a value that is greater than “< 1.” The most appropriate approach for addressing multiple censoring limits depends upon the nature of the parametric test being used. One approach consists of setting all of the non-detects to the largest censoring limit and treating these as tied values. Detected values less than the largest censoring limit (i.e., detection limit) must also be censored to the highest detection limit and treated as ties. This approach is not optimal because information is lost when all of the results are censored to the highest detection limit. However, the approach is statistically valid, simple to implement, and could be adequate for a large data set. It should also be noted that, rather than treating the NDs as ties, it is a common practice to rank the NDs according to their numerical estimates (EPA 68-W0-0025). Although this approach is used in this document (to be consistent with EPA guidance), it is not necessary appropriate. It is preferrable to use nonparametric statistical methods designed to account for multiple censoring limits such as the “Gehan test” and “generalized Wilcoxon test” described in Helsel (2005). If these methods are not available, consider assigning the largest censoring limit to all the non-detects.

H-2.6. Overview Summary. Some general guidelines are presented in Table H-1 based on the percentage of NDs. Substitution methods can potentially be used when less than 15% of the data are NDs. However, they are the preferred approach because the surrogate values that are substituted for the non-detects tend to distort the data sets to some degree. Adjustment or nonparametric methods should be considered, especially when more than 15% of the results are censored. If more than 50% of the data set’s concentrations are NDs, it is recommended that nonparametric methods be used instead of adjustment methods.

H-2.6.1. OSWER 9285.7-41/EPA 540-R-01-003 recommends a substitution method for censored results that is not recommended herein. The EPA suggests that a proxy value for NDs, based on one-half the censoring limit or on a random value between zero and the censoring limit, be used. According to the document, the censoring limit should be equal to the “sample-specific quantitation limit” and the method may be used so long as fewer than

50% of the data set's concentrations are NDs. However, it is recommended that proxy values not be used, especially when more than 15% of the results are reported as NDs. Using proxy values can bias the results of the statistical evaluations. The data user should verify that the "sample-specific quantitation limit" (SQL) is an appropriate censoring limit and adequately addresses false negatives as discussed in Appendix G. False negatives will not be minimized at the SQL when this limit is essentially a sample-specific MDL.

Table H-1.
Guidelines for Analyzing Data with NDs

Percentage of NDs	Paragraph	Proxy Definition/Statistical Analysis Method
< 15	H-3	Replace NDs with one-half censoring limit or a very small number
15–50	H-4	Trimmed mean, Cohen's or Atchison's adjustment, Winsorized mean, and standard deviation or non-parametric methods
> 50 – 90	H-5	Use tests for proportions
> 90	H-6	Use tests based on Poisson distribution

H-2.6.2. Although guidelines in Table H-1 are usually adequate, they should be implemented cautiously. Professional judgment is critical. In particular, the use of proxy values for a substitution approach should be evaluated in terms of the data quality objectives of the project. If the censoring limits are greater than or near project decision levels, then this approach may not be appropriate.

H-2.6.3. In Table H-1, all of the suggested procedures for analyzing data with NDs depend on the percentage of data below the censoring limit. For relatively small amounts below the censoring limit, replacing the NDs with a small number and proceeding with the usual analysis may be satisfactory. For moderate amounts of data below the censoring limit, a more detailed adjustment is appropriate. In situations where relatively large amounts of data below the censoring limit exist, one may need only to consider whether a certain proportion of the samples display values greater than some threshold values. The interpretation of small, moderate, and large amounts of data below the censoring limit is subjective. Table H-1 provides guideline percentages to assist the user in evaluating their particular situation; however, it should be recognized that these percentages are not rigid rules, but should be based on judgment.

H-2.6.4. In addition to the percentage of samples below the censoring limit, sample size influences which procedures should be used to evaluate the data. For example, the case where the result for 1 sample out of 4 is not detected should be treated differently from the case where the results for 25 samples out of 100 are not detected. It is recommended that the data analyst consult a statistician for the most appropriate way to evaluate data containing values below the detection level.

H-2.6.5. The remaining portion of this Appendix describes in detail the various methods outlined above. Case studies and examples are also presented.

H-3. Substitution Methods for Less than 15% NDs. If small proportions, 15% or fewer, of the observations are NDs, these may be replaced with a small number, the DL, DL/2, or a random value between the DL and zero (see EPA 540-R-01-003). After the non-detected values have been given a proxy value, then the usual statistical analysis may be performed. If simple substitution of values below the DL 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.

H-3.1. As a simplified case study showing the magnitude of effect on simple statistics attributable to different proxy concentrations, consider the data in Table H-2 for sodium in surface soil at a site. This table presents summary statistics for sodium data when 3 results of the 21 samples analyzed are not detected and 8 types of proxy concentrations have been used to represent these non-detected results. These proxies are the DL, RL, ½DL, ½RL, and a random number selected in four different ways as described in the table.

H-3.2. Summary statistics, in particular the average and standard deviation, are affected by the choice of proxy concentration. Proxy concentrations were developed based on the sample-specific DL and the project RL to illustrate how they are affected by the limit used for estimation. In this case study, a concentration not detected is reported as < DL. The DL is more appropriate to use to estimate a proxy than the RL, because it is the closest value at which the non-detected concentration may have occurred. If a concentration was > DL, but still < RL, the concentration is reported as a detect. Hypothetically, had only the RL been available and no DL had been provided, an alternative method to determine a proxy concentration would be to select the lower of the RL and the minimum detected result. Then, the proxy value would be at least below all of the detected concentrations.

H-3.3. As a basis for comparison, the summary statistics were also calculated using only the positively detected results in column 1 of Table H-2. In this instance, it is expected that the calculated average concentration would be higher than the true average, and the calculated standard deviation would be lower than the true standard deviation. When the RL is used to create proxy values (columns 7, 9, and 11), the average is higher and the standard deviation is lower than the associated summary statistics when the DL is used (columns 6, 8, and 10). Of all the cases when the RL is used to create a proxy, the case when a random number between zero and the RL is used (column 11) tends to have estimates for the average and standard deviation that are similar to the cases when the DL is used. This may be related to the fact that when a simple substitution such as the RL or ½RL is used, the variability is reduced because the proxy concentrations do not account for the inherent variation among concentrations. Proxy values are consistently the same number, whereas a proxy value based on a random number varies. It is also interesting to note that, in general, the summary statistics are similar for cases using random numbers as proxy values, no matter if the proxy value was based on the DL, RL, or the lowest detected result.

H-4. Methods for 15 to 50% NDs. Adjustment methods for treating NDs are commonly applied when NDs compose 15 to 50% of the data set. These various methods have their strengths and weaknesses, and they are presented first. Cohen's method is probably the most frequently used. A brief outline of a non-parametric procedure follows the discussion of adjustment methods.

H-4.1. Cohen's Method. Cohen's method provides adjusted estimates of the sample mean and standard deviation that accounts for data below the detection level when data are normally distributed. The adjusted mean and standard deviation can then be used in the parametric test described in Appendix L (EPA/240/B-026/003, QA/G-9S). This method requires knowing the censoring level, the percent of NDs, and either the arithmetic mean and standard deviation of the data (if the data are normally distributed) or the arithmetic mean and standard deviation of the log-transformed data (if the data are log-normally distributed). The data must also be evaluated for normality (Appendix F). For Cohen's method, the distribution is tested on the entire data set: positive detections and censored data. If the distribution testing fails to be normal or lognormal, Atchison's method (described below) may be more appropriate. Once the data distribution has been determined to be normal, the proxy concentrations themselves are essentially irrelevant when computing the adjusted mean and standard deviation.

H-4.1.1. Cohen's adjustment is a theoretically attractive method for handling cases with between 15 and 50% NDs. Conceptually, the method considers the detected results to be the top *X*% of an assumed distribution (normal, lognormal). The mean and standard deviation are then computed by filling in the bottom *Y*% of the assumed distribution (i.e., by assuming that the NDs represent the lower tail of the assumed distribution). These are referred to as the adjusted mean and adjusted standard deviation. This method appears to be a reasonable method for handling NDs and is attractive because it does not require the use of proxy concentrations (after normality has been determined).

H-4.1.2. There are, however, several practical difficulties encountered when applying this method, as follows.

H-4.1.2.1. Because there are no tests for how reliably the top *X*% of the data represent the top *X*% of a normal or lognormal distribution, there is a high degree of reliance on subjective judgment in selecting the appropriate distribution. So, the dilemma remains whether it is more appropriate to determine the distribution based on just the detected values or whether it is more appropriate to determine the distribution based on detects and proxy concentrations representing the NDs.

H-4.1.2.2. With Cohen's method, the sample size is effectively reduced because estimates are based only on the detected results. Estimates that are based on a small number of results are highly sensitive to the degree of uncertainty. This is particularly true when a lognormal distribution is assumed, and there is a high proportion of ND results. This leads to poorer estimates of the standard deviation, which can substantially impact calculations.

H-4.1.2.3. The method assumes that a single censoring level applies to all ND results. This is not always true (for example, if some NDs are for diluted samples and others are not), and the selection of the censoring level used in the calculations can have a substantial effect on the outcome.

H-4.1.3. Because this method requires knowing the censoring level, which is sometimes not reported and sometimes differs from one sample to another when it is reported, the following recommendations should be followed when Cohen's method is used.

H-4.1.3.1. If the censoring level (DL) is reported, and is the same for all non-detected results, use this value.

H-4.1.3.2. If the censoring level (DL) is reported, but is not consistent across all non-detected results, it is preferable to use the minimum censoring level among the NDs if there is justification to do so.

H-4.1.3.3. If the censoring level is not reported, then the values used to compute the proxy concentrations is the lesser of the RL and the minimum detected result.

H-4.1.3.4. If the RL and minimum detected results are the same for all non-detects, use that value.

H-4.1.3.5. If the RL values are not consistent across all non-detected results, use the minimum value among the NDs if there is justification to do so.

H-4.1.4. Because of the unrealistically elevated summary statistics that result when Cohen's method is applied, this method should be used with caution. Using Cohen's method is not recommended in more complicated evaluations, such as those required for an analysis of variance. Despite the limitations, there may be specific instances where it is applicable. In these cases, the results should be examined carefully to ensure that the conclusions are reasonable. The computational details of Cohen's method are presented in Paragraph H-4.2, and an example is given in Paragraph H-4.3.

H-4.2. Directions for Cohen's Method. Let $x_1, x_2 \ldots x_m, \ldots, x_n$ represent the n data points with the first m values representing the data points above the DL. Thus, there are $(n - m)$ data points below the DL.

H-4.2.1. Verify the distribution of the data to determine if they follow a normal or lognormal distribution (Appendix F). If they follow a normal distribution, then the raw data should be used for the following calculations. If the data follow a lognormal distribution, then the log-transformed data should be used for the following calculations.

H-4.2.1.1. Compute the sample mean $\bar{x}_d$ from the data above the DL:

$$\bar{x}_d = \frac{1}{m}\sum_{i=1}^{m} x_i \, .$$

H-4.2.1.2. Compute the sample variance s_d^2 from the data above the DL:

$$s_d^2 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{m} \, .$$

H-4.2.1.3. Compute

$$h = \frac{n-m}{n}$$

and

$$\gamma = \frac{s_d^2}{(\bar{x} - DL)^2} \, .$$

Table H-2.
Case Study, Sodium in Surface Soil: Summary Statistics Using Various Substitution Methods for Proxy Values

		1	2	3	4	5	6	7	8	9	10	11
Sampling Event	Sample	Result (mg/kg)	DL (mg/kg)	RL (mg/kg)	1/2DL as Proxy	1/2RL as Proxy	DL as Proxy	RL as Proxy	Random Number between 0 and DL as Proxy	Random Number between 0 and RL as Proxy	Random Number between 0 and lower of min. result and DL as Proxy	Random Number between 0 and lower of min. result and RL as Proxy
A	SS-010	ND	50	500	25	250	50	500	18.4	68.4	4.4	7.3
A	SS-020	ND	50	500	25	250	50	500	24.5	272.0	38.1	56.3
A	SS-030	1710										
A	SS-040	1860										
A	SS-050	2150										
A	SS-060	ND	50	500	25	250	50	500	13.9	47.8	48.1	28.6
B	SB01	750										
B	SB02	2430										
B	SB03	1160										
B	SB04	66										
B	SB05	140				Positive detections for columns 2 through 11 are the same as reported in column 1. They are omitted to highlight the DL, RL, and various proxy values.						
B	SB06	89										
B	SB07	120										
B	SB08	60										
B	SB09	107										
B	SB10	170										
B	SB11	180										
B	SB12	310										
B	SB13	71										
B	SB14	88										
B	SB15	61										
Summary Statistics												

	1	2	3	4	5	6	7	8	9	10	11
25th Percentile	88.25			66	89	66	89	66	71	66	66
Median	155			120	180	120	180	120	140	120	120
75th Percentile	1057.5			750	750	750	750	750	750	750	750
Average	640.1			552.2	584.4	555.8	620.1	551.4	567.2	553.0	553.1
Standard Deviation	828.9			795.4	776.9	792.9	765.8	796.0	786.8	794.9	794.8

DL = Sample-specific Detection Limit
RL = Project Reporting Limit

H-4.2.1.4. Use h and γ in Table B-3 of Appendix B 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 (Paragraph H-4.4) to estimate $\hat{\lambda}$.

H-4.2.1.5. Estimate the corrected sample mean, $\bar{x}$, and sample variance, s^2, to account for the data below the DL as follows:

$$\bar{x} = \bar{x}_d - \hat{\lambda}(\bar{x}_d - DL)$$

$$s^2 = s_d^2 + \hat{\lambda}(\bar{x}_d - DL)^2.$$

H-4.2.2. If these estimates are based on the log-transformed data, then they can be transformed back to the original units to estimate the mean and variance of the lognormal distribution. For example, if n is large, the mean and variance (of the untransformed data set) can be calculated as follows:

$$\bar{x}_{Ln} = \exp\left(\bar{x} + \frac{s^2}{2}\right) \text{ and } s_{Ln}^2 = \bar{x}^2\left[\exp\left(s^2\right) - 1\right] .$$

H-4.3. Example—Application of Cohen's Method. Of the groundwater analyses for benzene at Site A in monitoring well MW03, 10 of the 15 sample results are positive detections and 5 of the 15 sample results are NDs. Table H-3 presents the benzene concentrations, the DL, and natural log of the concentrations.

H-4.3.1. The total number of samples $n = 15$, the number of detects $m = 10$, and the number of non-detects $n - m = 5$.

H-4.3.2. The distribution of the positive detections was determined by the Shapiro-Wilk test (Appendix F) to be lognormal. The distribution of the entire data set, including NDs set to the proxy concentration equal to the DL, was also tested and evidence of a lognormal distribution was found. Thus, the Cohen's adjustment may be used.

$$\bar{x}_d = \frac{1}{10}\sum_{i=1}^{10} Ln(x_i) = -0.1650.$$

$$s_d^2 = \frac{\sum_{i=1}^{10}(Ln(x_i) - \bar{x}_d)^2}{10} = 1.683.$$

$$h = \frac{n-m}{n} = \frac{5}{15} = 0.3333 \ .$$

H-4.3.3. The DLs vary for the NDs. The lowest DL associated with the NDs will be used in these calculations. So, $DL = \text{Ln}(0.0375) = -3.283$.

$$\gamma = \frac{1.683}{(-0.1650-(-3.283))^2} = 0.1731 \ .$$

Table H-3.
Benzene Concentrations, the DL, and Natural Log of the Concentrations

Sampling Event	Result, *X* (µg/L)	DL (µg/L)	*Ln*(*X*) (*Ln*[µg/L])
29-Jan-98	ND	0.0605	ND
18-Apr-98	1.78	0.0375	0.5766
15-Jul-98	ND	0.0375	ND
18-Oct-98	2.31	0.0375	0.8372
18-Apr-99	7.24	0.0469	1.980
18-Jul-99	1.85	0.0759	0.6152
20-Oct-99	0.308	0.0759	–1.178
1-Apr-00	2	0.0504	0.6931
17-Jul-00	0.143	0.0353	–1.945
16-Oct-00	0.235	0.0353	–1.448
17-Jan-01	ND	0.0641	ND
4-May-01	0.759	0.0401	–0.2758
28-Jul-01	0.222	0.0401	–1.505
5-Nov-01	ND	0.0465	ND
31-Jan-02	ND	0.0465	ND

H-4.3.4. Using $h = 0.3333$ and $\gamma = 0.1731$ in Table B-3 of Appendix B and double linear interpolation (see Paragraph H-4.4 for details), $\hat{\lambda} = 0.5020$,

$$\overline{x} = -0.1650 - \left\{0.5020 \times \left[-0.1650-(-3.283)\right]\right\} = -1.730$$

and

$$s^2 = 1.683 + \left\{0.5020 \times \left[-0.1650-(-3.283)\right]^2\right\} = 6.563 .$$

H-4.3.5. Though *n* is relatively small, for the purposes of illustration, the corrected sample mean and variance for the lognormal distribution (based on the original units) are calculated as discussed in Paragraph H-4.2.

$$\bar{x}_{Ln} = \exp\left(-1.730 + \frac{6.563}{2}\right) = 4.719$$

and

$$s_{Ln}^{2} = (-1.730)^{2}\left[\exp(6.563) - 1\right] = 2117.$$

H-4.4. Double Linear Interpolation. The details of the double linear interpolation are provided to assist in the use of Table B-3 of Appendix B. Suppose the desired value corresponds to $\gamma = 0.1731$ and $h = 0.3333$ from Paragraph H-4.3. The values $\hat{\lambda}$ from Table B-3 for interpolation are:

γ	$H = 0.30$	$h = 0.35$
0.15	0.4330	0.5296
0.20	0.4422	0.5403

H-4.4.1. There are 0.05 units between 0.30 and 0.35 on the h scale, and 0.0333 units between 0.30 and 0.3333. Therefore, the value of interest lies (0.0333/0.05)1000% = 66.6% of the distance along the interval between 0.30 and 0.35. To linearly interpolate between tabulated values on the h axis for $\gamma = 0.15$, the range between the values must be calculated, 0.5296 – 0.4330 = 0.0966; the value that is 66.6% of the distance along the range must be computed, 0.0966 × 0.666 = 0.06434; and then that value must be added to the lower point on the tabulated values, 0.4330 + 0.06434 = 0.4973. Similarly for $\gamma = 0.20$, 0.5403 – 0.4422 = 0.0981, 0.0981 × 0.666 = 0.06533, and 0.4422 + 0.06544 = 0.5075. So,

γ	$h = 0.30$	$h = 0.3333$	$h = 0.35$
0.15	0.4330	0.4973	0.5296
0.20	0.4422	0.5075	0.5403

H-4.4.2. On the γ-axis there are 0.0231 units between 0.15 and 0.1731, and there are 0.05 units between 0.15 and 0.20. The value of interest (0.1731) lies (0.0231/0.05)100% = 46.2% of the distance along the interval between 0.15 and 0.20, so 0.5075 – 0.4973 = 0.0102, 0.0102 × 0.462 = 0.004712. Therefore, $\hat{\lambda} = 0.4973 + 0.004712 = 0.5020$.

H-4.5. Atchison’s Method. Previous adjustments to the mean and variance assumed that the data values really were present, but could not be recorded or seen as they were below the DL. In other words, if the DL had been substantially lower, the data values would have been recorded. There are cases, however, where the data values are below the DL because they are actually not present, the contaminant or chemical of concern being entirely absent. The investigator may have reason to believe that the contaminant is absent, but is unable to prove it is below the analytical DLs. Such data sets are actually a mixture—partly the assumed distribution (for example, a normal distribution) and partly a number of real zero values. Atchison's method is used in this situation to adjust the mean and variance for the zero values. It should also be noted that Atchison's method differs from Cohen's method, in that, for Atchison's method, a normality test is performed for the detected results only.

H-4.5.1. Atchison's method for adjusting the mean and variance of the values above the DL works quite well provided the percentage of NDs is between 15 and 50% of the total number of values. Care must be taken when using Atchison's adjustment because the mean is reduced and variance increased. With such an effect, it may become very difficult to use the adjusted data for tests of hypotheses or for predictive purposes.

H-4.5.2. As a diagnostic tool, Atchison's adjustment can lead to an evaluation of the data to determine if two populations are being sampled simultaneously: one population being represented by a normal distribution, the other being simply blanks. In some circumstances, such as investigating a hazardous site, it may be possible to relate the position of the sample through a posting plot and determine if the target population has not been adequately stratified. Directions for Atchison's method are contained in Paragraph H-4.6, and an example is contained in Paragraph H-4.7.

H-4.6. <u>Directions for Atchison's Method to Adjust Means and Variances</u>. Let $x_1, x_2, \ldots, x_m, \ldots, x_n$ represent the data points where the first *m* values are above the DL and the remaining (*n* – *m*) data points are below the DL.

H-4.6.1. Using the data above the detection level, verify this subset of data follows a normal distribution.

H-4.6.2. Using the data above the detection level, compute the sample mean,

$$\bar{x}_d = \frac{1}{m}\sum_{i=1}^{m} x_i$$

and the sample variance,

$$s_d^2 = \frac{\sum_{i=1}^{m}(x_i - \bar{x}_d)^2}{m-1}.$$

H-4.6.3. Estimate the corrected sample mean,

$$\bar{x} = \frac{m}{n}\bar{x}_d$$

and the sample variance,

$$s^2 = \frac{m-1}{n-1} s_d^2 + \frac{m(n-m)}{n(n-1)} \bar{x}_d^2 .$$

H-4.7. Example for Atchison's Method to Adjust Means and Variances. Atchison's method will be used to adjust the mean and standard deviation of the groundwater concentrations for benzene at Site A and well MW03, presented in Paragraph H-4.3.

H-4.7.1. So, $n = 15$, $m = 10$, and $n - m = 5$.

H-4.7.2. According to Paragraph H-4.3, the detected results from this data set follow a lognormal distribution; so, the log-transformed data will be used to adjust the mean and variance. The sample mean and variance based on just the data above the detection level are

$$\bar{x}_d = -0.1650$$

and

$$s_d^2 = 1.683 .$$

H-4.7.3. The corrected sample mean and variance (in the log-scale) are:

$$\bar{x} = \frac{10}{15} \times (-0.1650) = -0.1100$$

$$s^2 = \frac{10-1}{15-1} \times 1.683 + \frac{10(5)}{15(15-1)} \times (-0.1650)^2 = 1.088 \ .$$

H-4.8. Selecting Between Atchison's Method or Cohen's Method. To determine if a data set is better adjusted by Cohen's method or Atchison's method, a simple graphical procedure using a normal probability plot can be used. Directions for this procedure are given in Paragraph H-4.9, and an example is contained in Paragraph H-4.10.

H-4.9. Directions for Selecting Between Atchison's Method or Cohen's Method. Let $x_1, x_2, \ldots, x_m, \ldots, x_n$ represent the data points with the first *m* values above the DL and the remaining *n-m* data points below the DL.

H-4.9.1. Use Paragraph H-4.3 to construct a Normal Probability Plot using all the data, but only plot the values above the detection level. This is called the Censored Plot.

H-4.9.2. Use Paragraph H-4.3 to construct a Normal Probability Plot using only those values above the detection level. This is called the Detects Only Plot.

H-4.9.3. If the Censored Plot is more linear than the Detects Only Plot, use Cohen's method to estimate the sample mean and variance. If the Detects Only Plot is more linear than the Censored Plot, then use Atchison's method to estimate the sample mean and variance.

H-4.10. Example for Selecting Between Cohen's Method or Atchison's Method.

H-4.10.1. This comparison will be made with the groundwater concentrations for benzene at Site A and well MW03, based on the log-transformed data presented in Paragraph H-4.3.

H-4.10.2. Using Paragraph H-4.3, we constructed normal probability plots based on the log-transformed data, as the data seem to follow a lognormal distribution based on the Shapiro-Wilk test. The Figure H-1 shows these plots. The Censored Plot was developed with just the detected results. The Detects Only Plot was developed with all of the data (using the DL as a proxy value), but only the detected results were plotted. The Detects Only Plot appears to fit a line better than the Censored Plot, so Atchison's Method seems to be the more appropriate method to estimate the sample mean and variance.

H-4.11. Trimmed Mean. Trimming discards the data in the tails of a data set to develop an unbiased estimate of the population mean. This method is considered useful when the data set is generally symmetric, and there are concerns about outlier data that might be mistakes or otherwise unexplainable.

H-4.11.1. For environmental data, NDs usually occur in the left tail of the data, so trimming the data can be used to adjust the data set to account for NDs when estimating a mean. Developing a p100% trimmed mean involves trimming p100% of the data in both the lower and the upper tail. Note that p must be between 0 and 0.5 as 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 where n represents the original number of samples.

H-4.11.2. The proportion trimmed depends on the total sample size (n), as a reasonable number of samples must remain for analysis. For approximately symmetrical distributions, a 25% trimmed mean (the mid-mean) is a good estimator of the population mean. However, environmental data are often skewed (asymmetrical), and in these cases a 15% trimmed mean may be a better estimator of the population mean. It is also possible to trim the data only to replace the NDs. For example, if 3% of the data are below the DL, a 3% trimmed mean could be used to estimate the population mean. Directions for developing a trimmed mean are contained in Paragraph H-4.12, and an example is given in Paragraph H-4.13. A trimmed variance is rarely calculated and is of limited use.

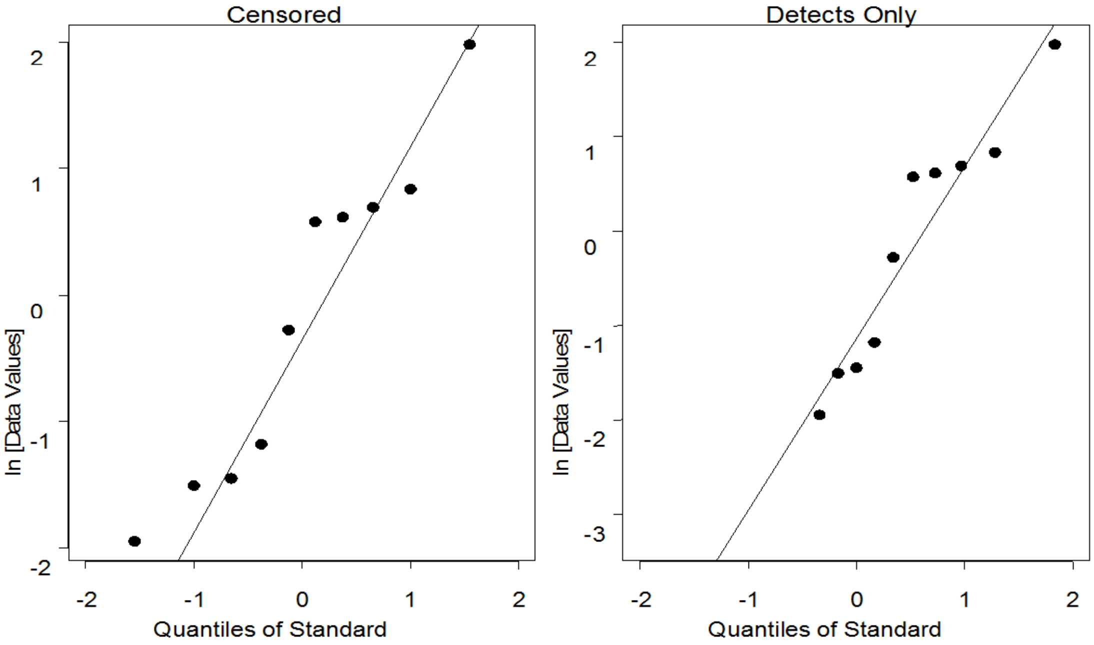


Figure H-1. Example of Selecting between Atchison's Method and Cohen's Method.

H-4.12. <u>Directions for Developing a Trimmed Mean</u>. Let $x_1, x_2, \ldots, x_n$ represent the *n* data points. To develop a *p*100% trimmed mean ($0 < p < 0.5$):

H-4.12.1. Let *j* represent the integer part of the product *np*. For example, if $p = 0.25$ and $n = 17$, $np = (0.25)(17) = 4.25$, so $j = 4$.

H-4.12.2. Delete the *j* smallest values of the data set and the *j* largest values of the data set.

H-4.12.3. Compute the arithmetic mean of the remaining $n - 2j$ values,

$$\bar{x} = \frac{1}{n-2j}\sum_{i=j+1}^{n-j-1} x_i \, .$$

This value is the estimate of the population mean.

H-4.13. <u>Example for Developing a Trimmed Mean</u>. For simplicity, a 100*p*% trimmed mean ($0 < p < 0.5$) will be estimated using the benzene data presented in the example in Paragraph H-4.3. As 5 out of 15 of the data are NDs, a 33.3% trimmed mean will be calculated.

$$n = 15$$

$$p = 0.333$$

$$np = 15 \times 0.333 = 5$$

$j = 5$ (the integer part of *np*) .

So, the 5 NDs and the 5 largest values of the data set will be removed, and the remaining samples will be used to estimate the average:

$$\bar{x} = \frac{1}{5}(0.308 + 0.143 + 0.235 + 0.759 + 0.222) = 0.3334.$$

H-4.14. 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, NDs usually occur in the left tail of the data. Winsorizing can be used to adjust the data set to account for NDs, and 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 Paragraph H-4.15, and an example is in Paragraph H-4.16.

H-4.15. Directions for Developing a Winsorized Mean and Standard Deviation. Let $x_1, x_2, \ldots, x_m \ldots, x_n$ represent the *n* data points and *m* represent the number of data points above the DL, and hence *n* – *m* below the DL.

H-4.15.1. List the data in order from smallest to largest, including NDs. Label these points $x_{(1)}$, $x_{(2)}$,..., $x_{(n)}$ such that $x_{(1)}$ is the smallest, $x_{(2)}$ is the second smallest, …., and $x_{(n)}$ is the largest.

H-4.15.2. Replace the *n* – *m* non-detects with $x_{(m+1)}$ and replace the *n* – *m* largest values with $x_{(n-m)}$.

H-4.15.3. Using the revised data set, compute the sample mean, $\bar{x}$, and the sample standard deviation, *s*:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} .$$

H-4.15.4. The Winsorized mean $\bar{x}_w$ is equal to $\bar{x}$. The Winsorized standard deviation is

$$s_w = \frac{s(n-1)}{2m-n-1}.$$

H-4.16. Example for Developing a Winsorized Mean and Standard Deviation. A Winsorized mean and standard deviation will be estimated using the groundwater concentrations for benzene at Site A and well MW03. Table H-4 presents these concentrations ordered from smallest to largest, where the NDs are considered the lowest concentrations. The five NDs are replaced by the smallest detected result (the 6th highest result) of 0.143, and the highest five detected results are replaced with the 10th highest result of 0.759.

$n = 15$, $m = 10$, $n - m = 5$.

Table H-4.
Groundwater Concentrations for Benzene at Site A for Well MW03

Sampling Event	Detected Result (µg/L)	DL (µg/L)	Revised Data (NDs replaced with smallest detected result)
15-Jul-98	ND	0.0375	0.143
05-Nov-01	ND	0.0465	0.143
31-Jan-02	ND	0.0465	0.143
29-Jan-98	ND	0.0605	0.143
17-Jan-01	ND	0.0641	0.143
17-Jul-00	0.143	0.0353	0.143
28-Jul-01	0.222	0.0401	0.222
16-Oct-00	0.235	0.0353	0.235
20-Oct-99	0.308	0.0759	0.308
04-May-01	0.759	0.0401	0.759
18-Apr-98	1.78	0.0375	0.759
18-Jul-99	1.85	0.0759	0.759
01-Apr-00	2.00	0.0504	0.759
18-Oct-98	2.31	0.0375	0.759
18-Apr-99	7.24	0.0469	0.759

H-4.16.1. Using the revised data set, we find the sample mean to be $\bar{x} = 0.4118$; this value is also the Winsorized mean. Using the revised data set, we find the sample standard deviation to be $s = 0.2970$.

H-4.16.2. The Winsorized standard deviation is

$$s_w = \frac{s\,(n-1)}{2m-n-1} = \frac{0.2970\,(15-1)}{(2\times 10)-15-1} = 1.0395.$$

H-4.17. Nonparametric Procedure. Another procedure that may be used, when the percent of NDs is between 15 and 50%, is a nonparametric analysis. First, all the data values need to be ordered and then replaced by their ranks. The NDs are then treated as tied values and replaced by their mid-ranks. The ranking procedure and adjustments for tied ranks are routinely performed for non-parametric tests such the Wilcoxon rank sum test.

H-5. 50 to 90% NDs. If more than 50% of the data are below the DL but at least 10% of the observations are quantified, tests of proportions may be used to test hypotheses using the data. 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 DL. For example, if 67% of the data are below the DL, consider switching the parameter of interest to the 75th percentile. Then, the test of proportion can be applied to test the hypothesis concerning the 75th percentile. It is important to note that tests of proportions may not be applicable for composite samples. In this case, the data analyst should consult a statistician before proceeding with analysis.

H-6. Greater than 90% NDs. The Poisson distribution can be used when 90% or more of the data is non-detected. In this instance, the detected results would be considered the “rare events” as modeled by the Poisson distribution. The Poisson model describes the behavior of a series of independent events over a large number of trials, where the probability of occurrence is low but stays constant from trial to trial. This model represents a counting process where each particle or molecule of contamination is counted separately but cumulatively, so that the counts for detected samples with high concentrations are larger than counts for samples with smaller concentrations. So, the Poisson model maintains the magnitude of detected concentrations. For example, a detected result with a concentration of 100 ppb would have a Poisson count of 100. Counts for non-detected results can be taken as zero or half the DL. The Poisson model is a distribution, like a normal distribution, that can be used to derive summary statistics such as prediction limits and tolerance limits. See Appendix E for a description of the Poisson distribution.

H-7. Recommendations.

H-7.1. If the degree of censoring (the percentage of data below the DL) is relatively low, reasonably good estimates of means, variances, and upper percentiles can be obtained. However, if the rate of censoring is very high (greater than 50%), then little can be done statistically except to focus on some upper quantile of the contaminant distribution, or on some proportion of measurements above a certain critical level that is at or above the censoring limit. Using nonparametric analyses is another approach for analyzing such data.

H-7.2. When the numerical standard is at or below one of the censoring levels and a one-sample test is used, the most useful statistical method is to test whether the proportion of a population is above (or below) the standard, or to test whether an upper quantile of the population distribution is above the numerical standard.

APPENDIX I

Identification and Handling of Outliers

I-1. Purpose.

I-1.1. 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 influence statistics if used in calculations, and statistical tests based on parametric methods are generally more sensitive than nonparametric methods to outliers. 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 and may indicate more variability in the population or a different underlying distribution for the population than what was initially assumed. For example, a point that appears as an outlier under the assumption that the underlying distribution is normal will not necessarily appear as an outlier if it were initially assumed that the distribution is lognormal. Not removing true outliers or removing false outliers can lead to a distortion of estimates of population parameters.

I-1.2. 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 a statistical outlier. These tests should only be used to identify data points that require further investigation. Tests alone cannot determine whether a statistical outlier should be discarded or corrected within a data set; this decision should be based on judgment and scientific reasoning. (See EPA 600/R-96/084, Gilbert, 1987, for further details on identifying and handling outliers.)

I-2. Methods. Five steps are involved in treating extreme values or outliers

1. Identify extreme values that may be potential outliers.

2. Apply a statistical test.

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

4. Conduct data analyses with and without statistical outliers.

5. Document the entire process.

Potential outliers can be identified through graphical representations. Graphs, such as the box- and-whisker plot, normal probability plot, and time plot, can be used to identify observations that are much larger or smaller than the rest of the data. (Appendix J presents these graphical tools.) If potential outliers are identified, the next step is to apply one of the statistical tests described below.

I-2.1. 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 extreme values that are much smaller or larger than the rest of the data. Because this test assumes that the data without the suspected outlier are normally distributed, it is necessary to test for normality in the data without the suspected outlier before applying Dixon's test. If the data are not normally distributed, a transformation that normalizes the data should be applied, or a different test should be used. Directions for the extreme value test are contained in Paragraph I-2.1.1 followed by an example in Paragraph I-2.1.2. Dixon's test should be used when only one outlier is suspected in the data. If more than one outlier is suspected, the extreme value test may lead to masking, in which two or more outliers close in value obscure one another. Therefore, if the analyst decides to use the extreme value test for multiple outliers, it should be applied to the least extreme value first; otherwise, Rosner's test should be used to test for multiple outliers. Rosner's test is discussed below.

I-2.1.1. Directions for the Extreme Value Test (Dixon's Test). Let $x_{(1)}, x_{(2)},...,x_{(n)}$ represent the data ordered from smallest to largest. Check that the data without the suspected outlier are normally distributed, using one of the methods in Appendix F.

I-2.1.1.1. If normality fails, transform the data or apply a different outlier test.

I-2.1.1.2. Case 1: $x_{(1)}$ is a potential outlier. Compute the test statistic C, where

$$C = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}} \text{ for } 3 \le n \le 7, \quad C = \frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}} \text{ for } 11 \le n \le 13,$$

$$C = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}} \text{ for } 8 \le n \le 10, \quad C = \frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}} \text{ for } 14 \le n \le 25.$$

I-2.1.1.3. If C exceeds the critical value from Table B-5 of Appendix B for the specified significance level α, $x_{(1)}$ is an outlier and should be further investigated.

I-2.1.1.4. Case 2: $x_{(n)}$ is a potential outlier. Compute the test statistic C, where

$$C = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \text{ for } 3 \le n \le 7, \quad C = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}} \text{ for } 11 \le n \le 13,$$

$$C = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} \text{ for } 8 \le n \le 10, \quad C = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}} \text{ for } 14 \le n \le 25.$$

I-2.1.1.5. If C exceeds the critical value from Table B-5 of Appendix B for the specified significance level α, $x_{(n)}$ is an outlier and should be further investigated.

I-2.1.2. Example for the Extreme Value Test (Dixon's Test). Consider the following subsurface background chromium data in order of magnitude from smallest to largest: 3.84, 4.26, 4.53, 4.60, 5.28, 5.29, 5.74, 5.86 (in mg/kg). Suppose there was an additional sample with a result of 10 mg/kg. As this additional sample is much larger than the other values, it is suspected that this point might be an outlier. The required level of significance for an outlier is 5%.

I-2.1.2.1. Testing the data for normality using the Shapiro-Wilk test (without the extreme value) indicated that the data were normal. Therefore, the extreme value test may be used to determine if the largest data value is an outlier.

$$C = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} = \frac{10.0 - 5.86}{10.0 - 4.26} = 0.72 \ .$$

I-2.1.2.2. Because $C = 0.72 > 0.512$ (from Table B-5 of Appendix B with $n = 9$ and $\alpha = 0.05$), there is evidence that $x_{(n)}$ is an outlier at a 5% significance level and should be further investigated.

I-2.2. Discordance Test. The discordance test can be used to test if one extreme value is an outlier. This test considers two cases: i) where the extreme value (potential outlier) is the smallest value of the data set; and ii) 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 the discordance test. If the data are not normally distributed, 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, so the test for normality should be performed without the suspected outlier. Directions and an example of the discordance test are contained in Paragraphs I-2.2.1 and I-2.2.2, respectively.

I-2.2.1. Directions for the Discordance Test. Let $x_{(1)}, x_{(2)},...,x_{(n)}$ 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 Appendix F, Paragraph F-11. If normality fails, transform the data or apply a different outlier test.

I-2.2.1.1. Compute the sample mean, $\bar{x}$, and the sample standard deviation, s, without the suspected outlier. If the minimum value $x_{(1)}$ is a suspected outlier, compute the test statistic

$$D = \frac{\bar{x} - x_{(1)}}{s}.$$

I-2.2.1.2. If D exceeds the critical value from Table B-4 of Appendix B, $x_{(1)}$ is an outlier and should be further investigated.

I-2.2.1.3. If the maximum value $x_{(n)}$ is a suspected outlier, compute the test statistic

$$D = \frac{x_n - \bar{x}}{s} .$$

I-2.2.1.4. If D exceeds the critical value from Table B-4 of Appendix B, $x_{(1)}$ is an outlier and should be further investigated.

I-2.2.2. Example for the Discordance Test. Consider the following subsurface background chromium data from smallest to largest: 3.84, 4.26, 4.53, 4.60, 5.28, 5.29, 5.74, 5.86 (in mg/kg). Suppose there was an additional sample with a result of 10 mg/kg. Because this additional sample is much larger than the other values, it is suspected that this point might be an outlier. The required level of significance for an outlier is 5%.

I-2.2.2.1. Testing the data for normality using the Shapiro-Wilk test (without the extreme value) indicated the data were normal. Therefore, the discordance test may be used to determine if the largest data value is an outlier.

$\bar{x} = 5.48$ mg/kg and $s = 1.82$ mg/kg without the suspected outlier.

I-2.2.2.2 Because the maximum value $x_{(n)}$ is a suspected outlier, do the following:

$$D = \frac{x_n - \bar{x}}{s} = \frac{10.0 - 5.48}{1.82} = 2.48.$$

I-2.2.2.3. Because $D = 2.48 > 2.110$ (from Table B-4 of Appendix B with $n = 9$ and $\alpha = 0.05$), there is evidence that $x_{(1)}$ is an outlier at a 5% significance level and should be further investigated.

I-2.3. Rosner's Test. Rosner developed a parametric test that 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, a test for normality should be performed before applying it. If the data are not normally distributed, 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, so the test for normality may be done without the suspected outlier. Directions for Rosner's test are contained in Paragraph I-2.3.2 and an example is contained in Paragraph I-2.3.3.

I-2.3.1. Caveats. Rosner's test is not as easy as the preceding tests to apply. To apply this test, first determine an upper limit r_0 for 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 B-18 of Appendix B, there are r_0 outliers. Otherwise, the test is performed again with the $r = r_0 - 1$ extreme values. This process is repeated until either Rosner's test statistic is greater than the critical value or $r = 0$.

I-2.3.2. Directions for Rosner's Test for Outliers. Let $x_{(1)}, x_{(2)},...,x_{(n)}$ 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 in Appendix F, Paragraph F-11.

I-2.3.2.1. Compute the sample mean, $\bar{x}$, and the sample standard deviation, s, for all of 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.

I-2.3.2.2 In summary, after the above process the analyst should have

$$\left[\bar{x}^{(0)},\, s^{(0)},\, y^{(0)}\right]\ ;\ \left[\bar{x}^{(1)},\, s^{(1)},\, y^{(1)}\right]\ ;\ \ldots,\ \left[\bar{x}^{(r_0-1)},\, s^{(r_o-1)},\, y^{(r_0-1)}\right]$$

where

$$\bar{x}^{(i)} = \frac{1}{n-i}\sum_{j=1}^{n-i} x_j\,, s^{(i)} = \left[\frac{1}{n-i}\sum_{j=1}^{n-1}\left(x_j - \bar{x}^{(i)}\right)^2\right]^{1/2}$$

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 were renumbered after each observation was deleted.)

I-2.3.2.3. To test if there are r outliers in the data, compute

$$R_r = \frac{\left|y^{(r-1)} - \bar{x}^{(r-1)}\right|}{s^{(r-1)}}\ .$$

Compare R_r to λ_r in Table B-18 of Appendix B. If, $R_r \geq \lambda_r$ conclude that there are r outliers. First, test if there are r_0 outliers (compare R_{r_0} to λ_{r_0}). If not, test if there are $r_0 - 1$ outliers (compare R_{r_0-1} to λ_{r_0-1}). If not, test if there are $r_0 - 2$ outliers, and continue until it is determined there are a certain number of outliers or no outliers at all.

I-2.3.3. Example for Rosner's Test for Outliers. Consider the following subsurface site copper data in order from smallest to largest: 1.99, 2.19, 2.34, 2.42, 2.45, 2.64, 2.70, 2.79, 2.82, 2.85, 2.86, 2.93, 3.10, 3.19, 3.21, 3.23, 3.25, 3.26, 3.28, 3.43, 3.55, 3.66, 3.71, 3.76, 3.83, 3.91, 3.92, 3.97, 3.98, 4.48, 5.0, 11.1, 11.6, 12.3, 32.1, 44.2.

I-2.3.3.1. By inspection, five potential outliers are suspected. Testing the data for normality using the Shapiro-Wilk test (without the extreme values) indicated that the data were normal. So Rosner's test for outliers may be used to determine if there are five or fewer outliers.

I-2.3.3.2. First the sample mean and sample standard deviation were computed for the entire data set, $\bar{x}^{(0)}$ and $s^{(0)}$. Subtraction showed that 44.20 was the farthest data point from $\bar{x}^{(0)}$, so $y^{(0)} = 44.20$. Then 44.20 was deleted from the data and the sample mean, $\bar{x}^{(1)}$, and the sample standard deviation, $s^{(1)}$, were computed. Subtraction showed that 32.10 was the farthest value from $\bar{x}^{(1)}$. This value was then dropped from the data and the process was repeated again on 12.30 and 11.60 to yield the values below.

i	$\bar{x}^{(i)}$	$s^{(i)}$	$y^{(i)}$
0	5.88	8.43	44.20
1	4.79	5.36	32.10
2	3.99	2.51	12.30
3	3.74	2.07	11.60
4	3.49	1.54	11.10

I-2.3.3.3. To apply Rosner's test, it is first necessary to test if there are five outliers (r = 5) by computing

$$R_5 = \frac{\left|y^{(4)} - \bar{x}^{(4)}\right|}{s^{(4)}} = \frac{|11.10 - 3.49|}{1.54} = \frac{7.61}{1.54} = 4.94$$

and comparing R_5 to λ_5 in Table B-18 of Appendix B with n = 36 and α = 0.05. Because $R_5 = 4.94 > \lambda_5 = 2.94$, there are five outliers in the data set.

I-2.3.3.4. Suppose $R_5 > \lambda_5 = 2.94$.

I-2.4. Walsh's Test. Walsh developed a nonparametric test to detect multiple outliers in a data set. This test requires a large sample size: n > 220 for a significance level of α = 0.05, and n > 60 for a significance level of $\alpha = 0.10$. However, as the test is nonparametric, it may be used whenever the data are not normally distributed. Directions for the Walsh test for large sample sizes are provided in Paragraph I-2.4.1, followed by an example in Paragraph I-2.4.2.

I-2.4.1. Directions for Walsh's Test for Large Sample Sizes. Let $x_{(1)}$, $x_{(2)}$,...,$x_{(n)}$ 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 $\alpha = 0.05$.

I-2.4.1.1. Identify the number of possible outliers, *r*. Note that *r* can equal 1.

I-2.4.1.2. Compute

$$c = \left[\sqrt{2n}\right],\ k = r + c,\ b^2 = 1/\alpha$$

and

$$a = \frac{1 + b\sqrt{(c - b^2)/(c-1)}}{c - b^2 - 1}$$

where [ ] indicates rounding the value up to the next largest integer (i.e., 3.24 becomes 4).

I-2.4.1.3. The *r* smallest points are outliers (with an α % level of significance) if

$$x_{(r)} - (1 + a)\,x_{(r+1)} + \ a\,x_{(k)} < 0.$$

I-2.4.1.4. The *r* largest points are outliers (with an α % level of significance) if

$$x_{(n+1-r)} - (1 + a)\,x_{(n-r)} + a\,x_{(n+1-k)} > 0.$$

I-2.4.1.5. If both of the inequalities are true, small and large outliers are indicated.

I-2.4.2. Example for Walsh's Test for Large Sample Sizes. Consider that the following surface soil lead data from Site 2 in order from smallest to largest: 11.7, 13.9, 14.4, 15.1, 17.2, 19.1, 19.3, 19.5, 19.6, 19.9, 20.8, 21.2, 21.8, 23.4, 24.2, 24.3, 25.8, 26.4, 27.4, 28.1, 29.1, 34.3, 35.3, 36, 37.9, 39.8, 43.8, 45.4, 51.4, 65.4, 74.4, 78.5, 87, 93.3, 105, 108, 120, 134, 135, 136, 143, 150, 178, 186, 194, 203, 214, 216, 232, 251, 263, 268, 277, 283, 300, 421, 446, 510, 811, 1260, 5320.

I-2.4.2.1. The possible outliers are 811, 1260, 5320. So *r* = 3.

$$c = \left[\sqrt{2n}\right] \ = \left[\sqrt{2 \times 63}\right] = [11.22] = 12$$

$$k = r + c = 3 + 12 = 15$$

$$b^2 = 1/\alpha = \frac{1}{0.10} = 10$$

$$a = \frac{1 + b\sqrt{(c-b^2)/(c-1)}}{c - b^2 - 1} = \frac{1 + 3.16\sqrt{(12-10)/(12-1)}}{12 - 10 - 1} = 2.347$$

$$x_{(n+1-r)} - (1+a)\, x_{(n-r)} + a\, x_{(n+1-k)} > 0$$

$$x_{(63+1-3)} - (1 + 2.347) x_{(63-3)} + 2.347 x_{(63+1-15)} > 0$$

$$x_{(61)} - (1 + 2.347) x_{(60)} + 2.347 x_{(49)} > 0$$

$$811 - (1 + 2.347)\, 510 + 2.347\, (214) > 0$$

$$-393.712 \not> 0\,.$$

I-2.4.2.2. Therefore the largest points, 811, 1260, 5320, are not outliers at $\alpha = 0.10$.

I-2.5. Fourth-Spread Outlier Test. A graphical qualitative method for identifying outliers entails creating box-and-whisker plots. Paragraph J-3 of Appendix J describes how to create such a plot. The process of identifying outliers by generating box-and-whisker plots is the same as identifying outliers using the “fourth-spread” outlier test (Hoaglin et al. 1983). The fourth-spread outlier test can identify one or more outliers from either end of the range of sample results.

I-2.5.1. A box-and-whisker plot identifies mild and extreme outliers. A mild outlier is a statistical outlier that is any result less than the difference of the 25th percentile and 1.5 times the inter-quartile range (IQR), or any result greater than the sum of the 75th percentile and 1.5 × IQR. An extreme outlier is a statistical outlier that is any result less than the difference of the 25th percentile and 3 × IQR, or any result greater than the sum of the 75th percentile and 3 × IQR. Extreme outliers are more severe than mild outliers and should be considered more influential.

I-2.5.2. The advantages of this test are that it does not have any sample size requirements and can identify one or more outliers. A disadvantage of the test is that no level of significance is placed on the decision to declare a result an outlier. However, it should be noted that, for a normally distributed variable X with a standard deviation of σ, 1.5 × IQR is approximately 2σ and there is slightly less than a 1% chance that points will be greater than $X_{0.75}$ + 1.5 × IQR or less than $X_{0.25}$ – 1.5 × IQR. Otherwise, the choice of 1.5 times the inter-quartile range is “somewhat arbitrary, but experience with many data sets indicates that this definition serves well in identifying values that may require special attention” (Hoaglin et al., 1983).

I-2.6. Multivariate Outliers. Multivariate analysis, such as factor analysis and principal components analysis, involves the statistical analysis of several variables simultaneously. Outliers in multivariate analysis are values that are extreme in relationship to one or more variables. As the number of variables increases, identifying potential outliers using graphical representations becomes more difficult. Special procedures are required to test for multivariate outliers. Details of these procedures are beyond the scope of this document, but are contained in statistical textbooks on multivariate analysis (see Gnanadesikan, 1997).

I-3. Retaining or Discarding Outliers. Once outliers are identified, the project team should review outliers and determine, case-by-case, if there is an explanation for each outlier. Furthermore, any suspect data point, whether identified as a statistical outlier or not, should be reviewed. Unexpected values, especially those identified as statistical outliers, should not be removed from any data evaluations unless a specific reason for the unexpected measurements can be determined.

I-3.1. If a data point is found to be an outlier, the analyst may: i) correct the data point; ii) discard the data point from analysis; or iii) use the data point in all analyses. Removing outliers should be based on scientific reasoning in addition to the results of the statistical test. An outlier should never be discarded based solely on a statistical test. Instead, the decision 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.

I-3.2. According to EPA 530-SW-89-026, a value may be corrected or dropped only if one can determine that an error has occurred. If an error can be identified, the correction should be made and the correct value used. Data points containing transcription errors should be corrected whether they are outliers or not. A value that is identified as incorrect may be deleted from the data set. Valid reasons for removing outliers or unexpected values include, for example, evidence they are the result of contaminated sampling equipment, laboratory errors, malfunctioning instrumentation, transcription errors, sampling of differing geological strata, or a non-typical sampling location taken for background. If a plausible reason cannot be found for removing an unexpected value or a statistical outlier, the result should be treated as a true but extreme value and retained in the data.

I-3.3. The spatial context of outliers or potential outliers should be considered. If outliers occur at different locations for different analytes and tend to be located close to low concentrations, then sporadic high concentrations are simply a feature of the area; there is no reason to treat the data differently as a result of their presence. If outliers tend to occur in the same location for different analytes and are found close to other locations with elevated concentrations, it may be appropriate to consider the elevated locations separately.

I-3.4. 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.

I-3.5. If any data points are found to be statistical outliers, this information should be documented along with the analysis of the data set, regardless of whether any data points are discarded. If no data points are discarded, the analyst should document that a process was implemented to identify any statistical outliers but none were found. If any data points are discarded, the analyst should 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. Such information is critical for effective peer review.

I-4. Applications. This Paragraph provides a case study regarding outliers and how conclusions are affected by including or excluding outliers. This case study focuses on identifying outliers in background data.

I-4.1. A background metals study was conducted to determine background concentrations that may be compared to site concentrations. Regulators were concerned with identifying outliers in the background data and removing them from the background data set, based upon the erroneous assumption that unusually high concentrations cannot represent background conditions and necessarily represent site-related contamination. All background data (by metal), were evaluated for outliers using two outlier tests—the discordance test and fourth-spread test. For this investigation, the regulator required that any result identified as a statistical outlier be removed from the background data set, which biased the background sample towards smaller values. This case study focuses on the evaluation of antimony in surface soil.

I-4.2. Table I-1 presents the 20 samples associated with antimony concentrations from the background surface soil. Generally, the concentrations were quite small, ranging from 0.182 to 0.398 mg/kg. Outlier tests were performed on the highest concentration, 0.398 at sample BACK-005-005, to see if this concentration could be considered a statistical outlier.

I-4.3. First, a box-and-whisker box plot was generated to visualize the data and to perform the fourth-spread test. As Figure I-1 presents with the box plot, the highest concentration is a mild outlier.

Table I-1.
Background Surface Soil Data for Antimony

Sample ID	Result (mg/kg)	Sample ID	Result (mg/kg)
BACK-001-0005	0.235	BACK-0011-0005	0.202
BACK-002-0005	0.285	BACK-0012-0005	0.27
BACK-003-0005	0.202	BACK-0013-0005	0.298
BACK-004-0005	0.22	BACK-0014-0005	0.209
BACK-005-0005	0.398	BACK-0015-0005	0.182
BACK-006-0005	0.279	BACK-0016-0005	0.233
BACK-007-0005	0.215	BACK-0017-0005	0.186
BACK-008-0005	0.25	BACK-0018-0005	0.267
BACK-009-0005	0.279	BACK-0019-0005	0.273
BACK-0010-0005	0.23	BACK-0020-0005	0.28

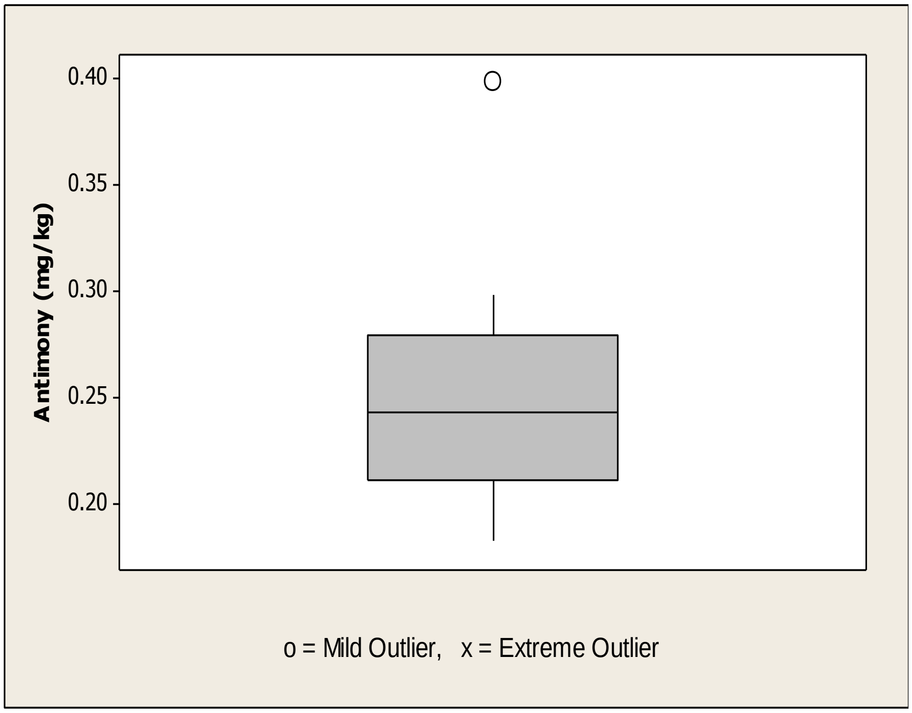


Figure I-1. Box-and-Whisker Plot for Antimony.

I-4.4. The discordance test was done to determine if the maximum result might be considered a statistical outlier. Results of the discordance test show the maximum result is an outlier, as seen below.

I-4.4.1. Normality Assumption. The Shapiro-Wilk test was performed on the raw data, without the maximum result. The test statistic for this test was 0.9319 and the *p* value associated with this test statistic was 0.1878. Based on 95% level of confidence, because 0.1878 > 0.05, there is evidence to suggest the data without the maximum result were normal. Therefore, doing the discordance test on the raw data was appropriate.

I-4.4.2. Test Statistic. $D = \frac{X_n - \bar{x}}{s} = \frac{0.398 - 0.2418}{0.0366} = 4.268$.

I-4.4.3. Critical Value. $2.557\ (\text{based on}\ \alpha = 0.05)$.

I-4.4.4. Conclusion. Because 4.268 > 2.557, there is evidence that the maximum result is an outlier.

I-4.5. As both outlier tests showed the maximum result is a statistical outlier, the maximum antimony result for surface soil was removed from the background data set at the request of the regulator even though the outlier appeared to be a valid result (i.e., it was not entered incorrectly or demonstrated to be the result of a non-complaint sampling or analytical procedure).

Table I-2.
Summary Statistics for Antimony Background Surface Soil Data

	All Samples	All but Max
n	20	19
Minimum (mg/kg)	0.182	0.182
Maximum (mg/kg)	0.398	0.298
Median (mg/kg)	0.2425	0.235
Mean (mg/kg	0.25	0.242
Standard Deviation (mg/kg)	0.04988	0.0366
95% UCL (mg/kg)	0.270	0.256
Distribution	Log-normal	Normal
***p* value for Shapiro-Wilk test for original data**	0.0369	0.1878
***p* value for Shapiro-Wilk test for log-transformed data**	0.3309	0.1667

I-4.6. From a statistical perspective, it was probably inappropriate to remove the maximum detected concentration as an outlier for the antimony data set. To illustrate this conjecture, separate lists of summary statistics are presented in Table I-2 for all 20 antimony results and for the 19 antimony results without the maximum concentration.

I-4.7. The most striking difference between the two data sets is their distribution. When all samples were evaluated, there was evidence that the data followed a lognormal distribution, but when all samples except the maximum were evaluated, there was evidence that the data followed both a normal and lognormal distribution. (A data point from a lognormal distribution can appear as an outlier when it is erroneously assumed that the data set is normally distributed.) However, for this particular data set, the removal of the outlier (0.398 mg/kg) did not significantly affect decision-making because all of the antimony concentrations were less than the state-specified risk-based decision level of 2.7 mg/kg. Furthermore, fortuitously, similar statistical results were obtained with and without the outlier. Although the maximum detected concentration was eliminated, the sample median and mean were not seriously affected, and the difference between maximum concentrations was less than an order of magnitude. However, under different circumstances (e.g., had the risk-based decision limit or the difference between the two highest values been larger), the comparisons between the site and background data sets could have been adversely affected (e.g., a "false positive" could have resulted). Data points should never be removed from any data set (background or otherwise) solely on the basis of an outlier test unless an independent weight of evidence indicates that the data points are not representative of the underlying population of interest.

I-5. Recommendations. If the data are normally distributed, Rosner's test is recommended when the sample size is greater than 25 and the extreme value test is recommended when the sample size is less than 25. If only one outlier is suspected, the discordance test 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, the analyst should apply a nonparametric test, such as the fourth-spread test, or Walsh's test. A summary of this information is contained in Table I-3. Recommendations on selecting a statistical test for outliers are listed.

Table I-3.
Recommendations for Selecting a Statistical Test for Outliers

Sample Size	Test	Assumes Normality	Multiple Outliers
$n \leq 25$	Extreme Value Test	Yes	No/Yes
$n \leq 50$	Discordance Test	Yes	No
$n \geq 25$	Rosner's Test	Yes	Yes
$n \geq 50$	Walsh's Test	No	Yes
Any sample size	Fourth-Spread Test	No	Yes

APPENDIX J

Graphical Tools

J-1. Introduction. Graphs are powerful data evaluation tools, providing a quick assessment of concentration ranges, extreme concentrations or data anomalies, and patterns and trends that may be unapparent otherwise. In exploratory data analysis, various graphical techniques are used initially to display the data so that users may determine what statistical evaluations will be used. Although a subjective assessment of a plot alone is often inadequate to make conclusions about the significance of a trend or association, plots support quantitative statistical tests.

J-1.1. This Appendix presents some common graphical methods for presenting environmental data in meaningful ways. These graphical methods are:

a. Histogram/Frequency Plots.

b. Box-and-Whiskers Plots.

c. Quantile Plots.

d. Normal Probability Plots (Quantile-Quantile Plots).

e. Empirical Quantile-Quantile Plots.

f. Plots for Temporal Data.

g. Plots for Spatial Data.

h. Plots for Two or More Variables.

i. Contouring Data.

J-1.2. Additional information on most of the plots presented here may be found in Mason et al. (1989). For temporal and spatial plots see EPA/240/B-026/003, QA/G-9S.

J-2. Histogram/Frequency Plots.

J-2.1. Introduction. Two of the oldest methods for summarizing data distributions are the frequency plot (Figure J-1) and histogram (Figure J-2). Both frequency plots and histograms divide the range of measured values of a variable into equal intervals, and use a bar graph to display the results. In a frequency plot, the height of each bar represents the number of observations within each interval. In a histogram, the height of each bar represents the percentage of observations within each interval.

J-2.1.1. 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 or number of observations within a group. In a histogram, the area within the bar represents the relative density of the data or percentage of observations within a group.

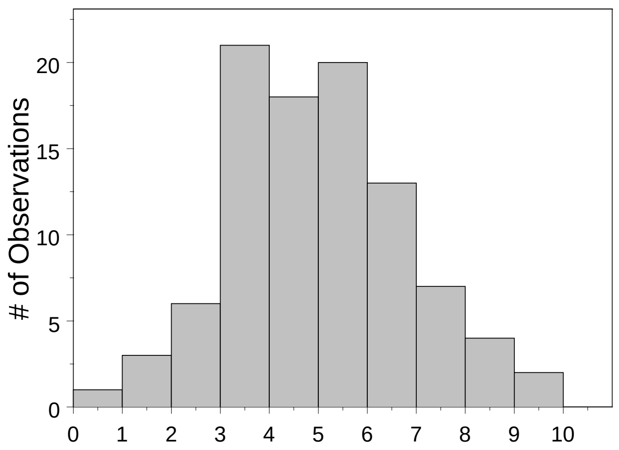


Figure J-1. Frequency Plot: Normal Data.

J-2.1.2. When plotting a histogram for a continuous variable (such as 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 (i.e., 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. For a continuous measurement variable, *X*, the histogram should approach the “true” probability distribution as the sample size increases and the width of the intervals decrease. For example, if the variable *X* is normally distributed, then the histogram will approach a Gaussian curve (see Appendix F). Figure J-1 plots 95 observations from a sample from a normal distribution with a mean of 5 and a standard deviation of 2. Notice how the histogram approximates a normal curve. Likewise, Figure J-2 plots 95 observations from a sample from a lognormal distribution with $\mu = 1$ and $\sigma = 1$.

J-2.1.3. Directions for generating a histogram and a frequency plot are presented in Paragraph J-2.2 and an example is contained in Paragraph J-2.3.

J-2.2. 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 do the following.

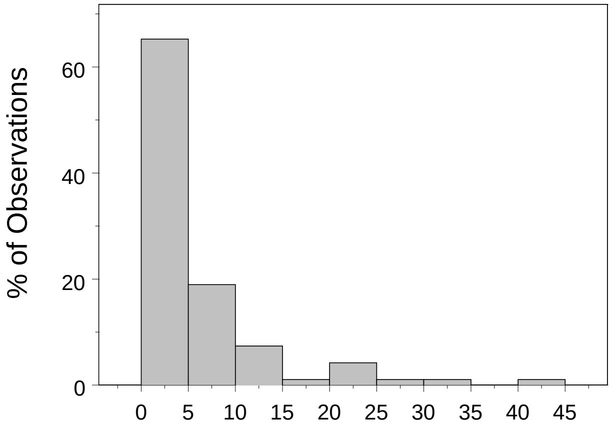


Figure J-2. Histogram: Lognormal Data.

J-2.2.1. 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.

J-2.2.2. 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.

J-2.2.3. 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 the percentage (or proportion) of results that fall within each interval on the
x-axis.

J-2.2.4. For a histogram, compute the percentage of observations within each interval by dividing the number of observations within each interval (Step J-2.2.3) by the total number of observations.

J-2.2.5. For a histogram, select a common unit that corresponds to the *x*-axis (Step J-2.2.1). Compute the number of common units in each interval and divide the percentage of observations within each interval (Step J-2.2.4) by this number. This step is only necessary when the intervals (Step J-2.2.1) are not of equal widths.

J-2.2.6. Using boxes, plot the intervals against the results of Step J-2.2.5 for a histogram or the intervals against the number of observations in an interval (Step J-2.2.2) for a frequency plot.

J-2.3. Example of a Histogram and a Frequency Plot. Consider the following results of benzene concentrations in groundwater (ppb): 0.0292, 0.0300, 0.0300, 0.0300, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0444, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0469, 0.0469, 0.0469, 0.0480, 0.0504, 0.0504, 0.0504, 0.0548, 0.0585, 0.0605, 0.0605, 0.0605, 0.0641, 0.0641, 0.0641, 0.0641, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0786, 0.0786, 0.0855, 0.0970, 0.0971, 0.1430, 0.2220, 0.2350, 0.3080, 0.4840, 0.6350, 0.7590, 0.8130, 1.1500, 1.7200, 1.7800, 1.8400, 1.8500, 1.9200, 2.0000, 2.0100, 2.1700, 2.1900, 2.3100, 2.4600, 2.6800, 2.7500, 2.9500, 3.4200, 3.4500, 3.7900, 4.3000, 5.4700, 5.7700, 5.8700, 6.1700, 6.9100, 7.2400, 7.5600, 8.3400, 8.6400, 9.3300, 11.000, 11.100, 12.200, 14.100, 17.000, 20.200, 21.800, 29.100, 36.700 and 44.500.

J-2.3.1. These data values span 0 to 50 ppb. Equally sized intervals of 5 ppb will be used: 0 to 5 ppb, 5 to 10 ppb, etc. The endpoint convention will be that values are placed in the highest interval containing the value. For example, a value of 5 ppb will be placed in the interval 5 to 10 ppb instead of 0 to 5 ppb. Table J-1 shows the number of observations within each interval defined here

J-2.3.2. The horizontal axis for the data is from 0 to 50. The vertical axis for the frequency plot is from 0 to 88 and the vertical axis for the histogram is from 0 to 81.5%.

J-2.3.3. There are 108 observations total, so the number of observations shown in the table will be divided by 108. The results are shown in the third column of the table.

J-2.3.4. A common unit for this data is 1 ppb. In each interval there are five common units so the percentage of observations (third column of the table) should be divided by 5 (fourth column).

J-2.3.5. The frequency plot (Figure J-3) and the histogram (Figure J-4) are shown below.

Table J-1.
Number of Observations within Each Interval

Interval (ppb)	Observations in Interval	Percent Observations in Interval	Percent Observations per ppb
0–5	88	81.5	16.3
5–10	10	9.26	1.85
10–15	4	3.70	0.74
15–20	1	0.926	0.185
20–25	2	1.85	0.370
25–30	1	0.926	0.185
30–35	0	0	0
35–40	1	0.926	0.185
40–45	1	0.926	0.185

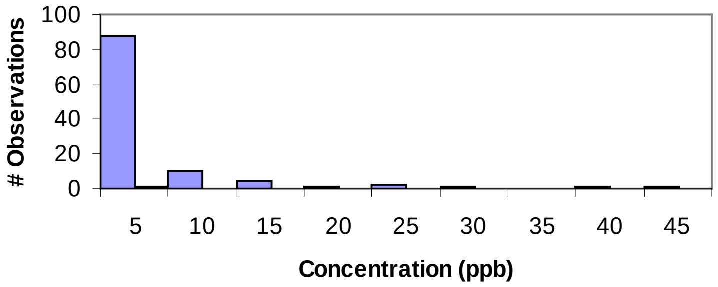


Figure J-3. Frequency Plot.

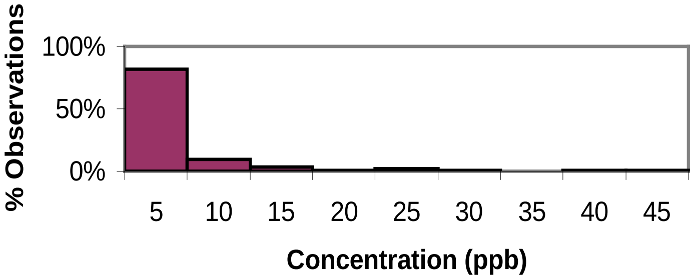


Figure J-4. Histogram.

J-3. Box-and-Whiskers Plots.

J-3.1. Introduction. A box-and-whiskers plot (or a box plot) is a schematic diagram useful for visualizing important statistical quantities such as the center, spread, and distribution of a data set.

J-3.1.1. A box-and-whiskers plot (Figure J-5) 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—the interquartile range (IQR) or the distance from the 25th to the 75th percentile—indicates the spread of the bulk of the data (the central 50%) while the length of the whiskers shows the extent of the tails in the distribution. The length of each whisker is 1.5 IQR (roughly equal to two standard deviations for a normal data set). 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 solid horizontal line through the box and the sample mean is displayed using a dotted horizontal line.

J-3.1.2. Box-and-whisker plots are useful for identifying possible outliers as they identify values that would be unusually large or small data if the data were assumed to be normally distributed. Any data points falling outside of the whiskers are displayed as “outliers” by an “o” or “x” on the plot. In particular, points falling 3.0 × IQR from the top or bottom of the box are “extreme outliers” displayed by an “x,” while points falling 1.5 × IQR (but within

3.0 × IQR) from the top or bottom of the box are “mild outliers” displayed by an “o.” For example, the box plot of the lognormal data in Figure J-5 contains three data values that are identified as unusual (two “mild outliers” and one “extreme outlier”) if the data were assumed to be from a normal distribution. Each of the features described in this paragraph has been labeled in Figure J-5 to help you identify the most important features of box plots.

J-3.1.3. A box-and-whiskers plot can also 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. For instance, the box plot of the normal data in Figure J-5 displays a highly symmetrical distribution of data. The mean and median are about the same, the 25th and 75th percentiles are about the same distance from the median, and the whiskers are roughly the same length. In contrast, the box plot of the lognormal data in Figure J-5 shows a noticeable positive skew. The mean is greater than the median, the upper whisker appears longer than the lower whisker, and several unusually large values are present on the upper end of the distribution. To see the variety in plots, the reader is urged to plot project-specific data.

J-3.1.4. Box-and-whiskers plots are extremely useful for visual comparisons of data from multiple sources when they are presented side-by-side. For example, separate box plots can be constructed for comparing background concentrations to site concentrations. This provides simultaneous comparison of the medians and IQRs of several sets of data. Another example where box plots can be useful is when trying to determine if an assumption of equal variances is valid, by qualitatively comparing the IQRs of two data sets (Appendix M). Directions for generating a box-and-whiskers plot are contained in Paragraph J-3.2 and an example follows in Paragraph J-3.3.

J-3.2. Directions for Generating a Box-and-Whiskers Plot.

J-3.2.1. 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 $-\frac{1}{2}W$ to $\frac{1}{2}W$.

J-3.2.2. Compute the upper quartile ($x_{0.75}$, the 75th percentile) and the lower quartile ($x_{0.25}$, the 25th percentile). Compute the sample mean and median. Compute the interquartile range (IQR). (Refer to Appendix D to do these computations, as necessary.)

J-3.2.3. Draw a box through points $(-\frac{1}{2}W, x_{0.75})$, $(-\frac{1}{2}W, x_{0.25})$, $(\frac{1}{2}W, x_{0.25})$, and $(\frac{1}{2}W, x_{0.75})$. Draw a line from $(\frac{1}{2}W, x_{0.50})$ to $(-\frac{1}{2}W, x_{0.50})$ and mark point $(0, \bar{x})$ with (+).

J-3.2.4. Compute the upper end of the top whisker by finding the largest data value X_L less than $x_{0.75} + 1.5 \times \text{IQR}$. Draw a line from $(0, x_{0.75})$ to $(0, x_\text{L})$. Compute the lower end of the bottom whisker by finding the smallest data value x_S greater than $x_{0.25} - 1.5 \times \text{IQR}$. Draw a line from $(0, x_{0.25})$ to $(0, x_S)$.

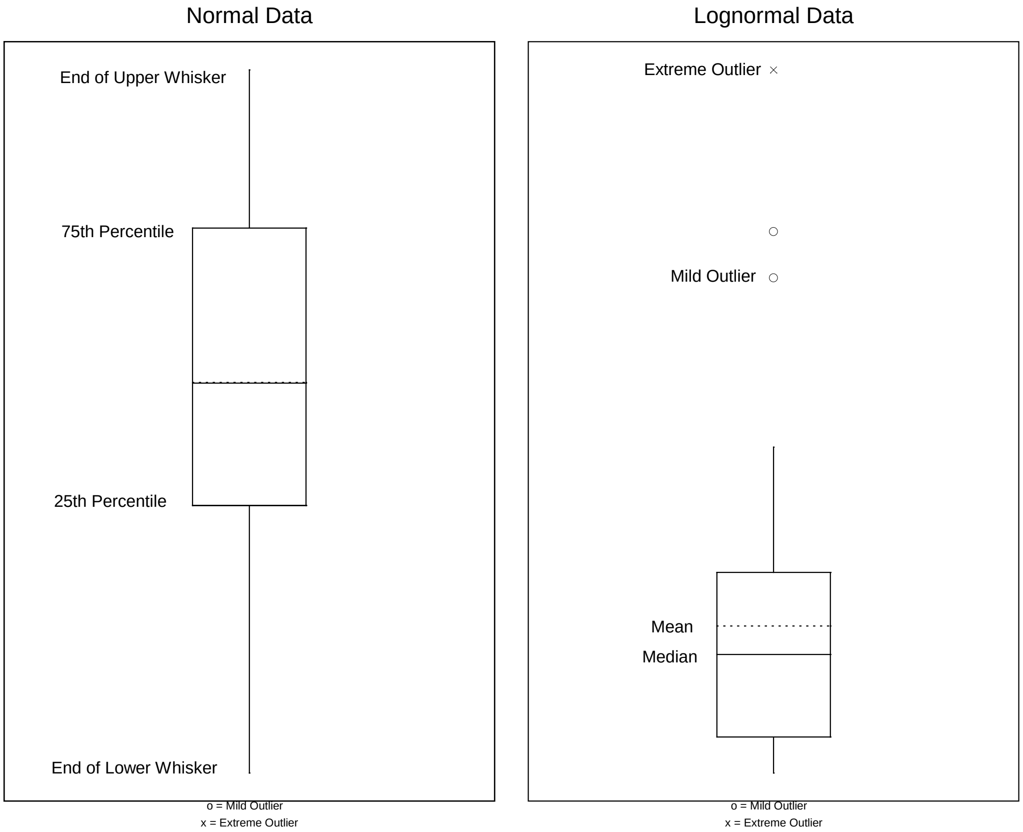


Figure J-5. Examples of Box-and-Whiskers Plots.

J-3.2.5. For all points $x_L < x_* < x_{0.75} + 3.0 \times \text{IQR}$, place an "o" at the point $(0, x_*)$. These points are considered mild outliers. For all points $x_{**} > x_{0.75} + 3.0 \times \text{IQR}$, place an "x" at the point $(0, x_{**})$. These points are considered extreme outliers. Likewise, for all points $x_{0.25} - 3.0 \times \text{IQR} < x_* < x_S$, place an "o" at the point $(0, x_*)$. Finally, for all points $x_{**} < x_{0.25} - 3.0 \times \text{IQR}$, place an "x" at the point $(0, x_{**})$.

J-3.3. Example of a Box-and-Whiskers Plot. Consider the following site data of chromium concentrations (mg/kg) in surface soil : 3.08, 3.35, 4.09, 4.13, 4.14, 4.36, 4.37, 4.42, 4.68, 4.76, 4.78, 4.82, 4.87, 4.89, 4.91, 4.94, 4.96, 4.96, 5.51, 6.4, 10.1, 10.3, 10.6 and 18.5

J-3.3.1. When generating the plot the width was set at a –0.25 to 0.25 horizontal range. Do not forget that the width is only a visualization tool and can be set to any value.

J-3.3.2. Compute the 75th percentile:

$p = 0.75$

$np = 24 \times .75 = 18$

$np = j + g = 18 + 0$

since g = 0

$$x_{0.75} = \frac{X_{(18)} + X_{(19)}}{2} = \frac{4.96 + 5.51}{2} = 5.235 \ .$$

J-3.3.3 Compute the 25th percentile:

$p = 0.25$

$np = 24 \times .25 = 6$

$np = j + g = 6 + 0,$

since g = 0

$$x_{0.25} = \frac{X_{(6)} + X_{(7)}}{2} = \frac{4.36 + 4.37}{2} = 4.365 \ .$$

Sample mean = 5.91, sample median = 4.845, interquartile range = $Q(.75) - Q(.25) = 0.87$.

J-3.3.4. Compute the upper end of the top whisker by finding the largest data value x_L less than

$x_{0.75} + 1.5 \times IQR = 5.235 + 1.5(0.87) = 6.54$.

So, x_L = 6.4. Draw a line from (0, 5.235) to (0, 6.4). Compute the lower end of the bottom whisker by finding the smallest data value x_S greater than

$x_{0.25} - 1.5 \times IQR = 4.365 - 1.5(0.87) = 3.06$.

So, x_S = 3.08. Draw a line from (0, 4.365) to (0, 3.08).

J-3.3.5. There are no points, x_*, greater than $x_L = 6.4$ but less than

$$x_{0.75} + 3.0 \times IQR = 5.235 + 3.0(0.87) = 7.845$$

so no points are considered mild outliers. For all points

$$x_{**} > x_{0.75} + 3.0 \times IQR = 7.845$$

place an “x” at the point (0, x_{**}). These points are considered extreme outliers. There are no points less than x_S = 3.08 so no points are drawn below the bottom whisker. Figure J-6 shows the box-and-whiskers plot.

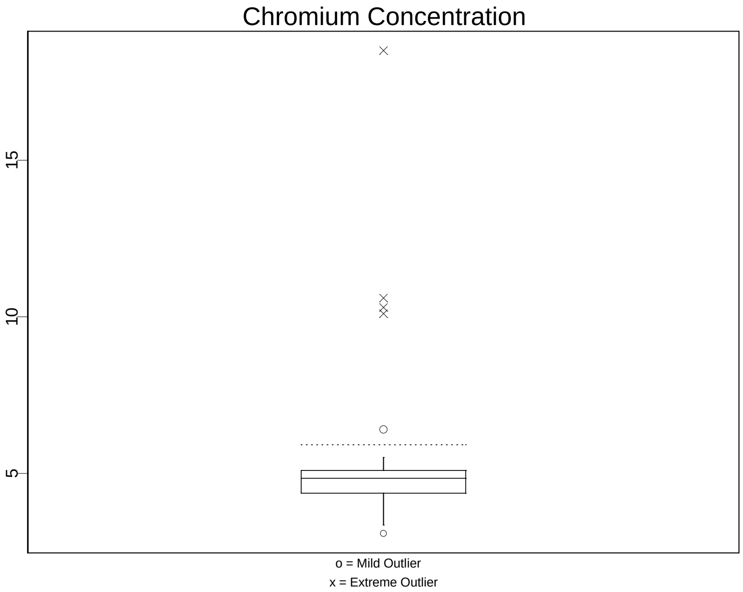


Figure J-6. Box-and-Whiskers Plot.

J-4. Quantile Plots.

J-4.1. Introduction. A quantile plot is a graph of the quantiles of data. It plots each point according to the fraction of the points it exceeds. It 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.

J-4.1.1. A quantile plot displays every data point ranging from the lowest value to the highest value; it is a graphical representation of the data instead of a summary of the data. The advantage of using a quantile plot is that the analyst does not have to make any arbitrary choices regarding the data to construct a quantile plot (such as selecting the cell sizes for a making a histogram).

J-4.1.2. The vertical axis of the quantile plot is the measured concentration, and the horizontal axis of the quantile plot is the percentile of the data distribution. Directions for developing a quantile plot are given in Paragraph J-4.2 and an example follows in Paragraph J-4.3.

J-4.1.3. A quantile plot can be used to read quantile information (the median, quartiles, and the interquartile range) because each data value is plotted against the percentage of the data with that value or less. In addition, the plot can be used to determine the density of the data points: Are all the data values close to the center with relatively few values in the tails or are there a large number of values in one tail with the rest evenly distributed? The density of the data is displayed through the slope of the graph. A flat slope indicates a large number of data values; the graph rises slowly. A steep slope indicates a small number of data values; the graph rises quickly. A quantile plot can be used to determine if the data are skewed or if they are symmetrical. Figure J-7 shows examples of three quantile plots. If the data are symmetrical, 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 similar to Figure J-7a. A quantile plot of data that are skewed to the right is steeper at the top right than the bottom left, as shown in Figure J-7b. A quantile plot of data that are skewed to the left increases sharply near the bottom left of the graph as shown in Figure J-7c.

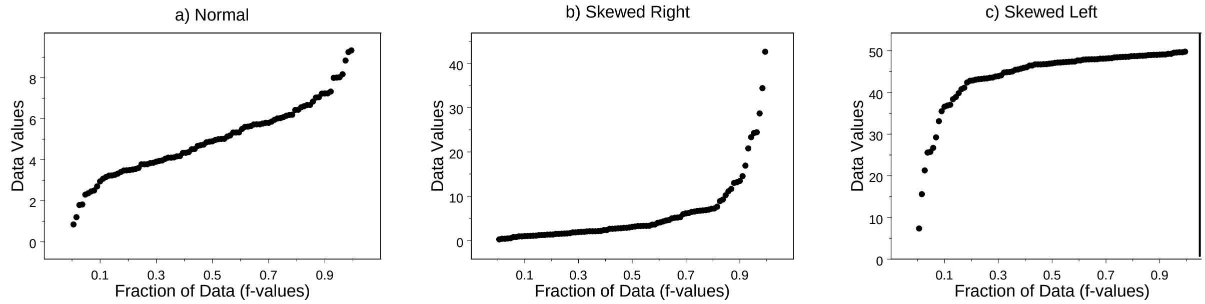


Figure J-7. Examples of Quantile Plots.

J-4.2. <u>Directions for Developing a Quantile Plot</u>. Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ represent the *n* data points ordered from least to greatest.

J-4.2.1. For each *i* from 1 to *n*, compute the fraction $f_i = (i - 0.5)/n$. The quantile plot is a plot of the pairs ($f_i, x_{(i)}$).

J-4.2.2. An example is given below in Paragraph J-4.3. (There are a number of ways to calculate the quantile $f_{i.}$ Software that performs quantile plots may not necessarily use the same formula presented in Paragraph J-4.2 to calculate the quantile. For example, for the Weibull method $f_i = i/(n+1)$.)

J-4.3. <u>Generating a Quantile Plot</u>. Consider the following 108 data points for benzene groundwater results in µg/L: 0.0292, 0.0300, 0.0300, 0.0300, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0444, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0469, 0.0469, 0.0469, 0.0480, 0.0504, 0.0504, 0.0504, 0.0548, 0.0585, 0.0605, 0.0605, 0.0605, 0.0641, 0.0641, 0.0641, 0.0641, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0786, 0.0786, 0.0855, 0.0970, 0.0971, 0.1430, 0.2220, 0.2350, 0.3080, 0.4840, 0.6350, 0.7590, 0.8130, 1.1500, 1.7200, 1.7800, 1.8400, 1.8500, 1.9200, 2.0000, 2.0100, 2.1700, 2.1900, 2.3100, 2.4600, 2.6800, 2.7500, 2.9500, 3.4200, 3.4500, 3.7900, 4.3000, 5.4700, 5.7700, 5.8700, 6.1700, 6.9100, 7.2400, 7.5600, 8.3400, 8.6400, 9.3300, 11.0000, 11.1000, 12.2000, 14.1000, 17.0000, 20.2000, 21.8000, 29.1000, 36.7000 and 44.5000.

J-4.3.1. The data, ordered from smallest to largest, $x_{(i)}$, are shown in the first column of Table J-2 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*.

Table J-2. Quantile Plot Data

$x_{(i)}$ (Mg/L)	***i***	f_i
0.0290	1	0.0046
0.0300	2	0.014
0.0300	3	0.023
.	.	.
.	.	.
.	.	.
29.100	106	0.9769
36.700	107	0.9861
44.500	108	0.9954

J-4.3.2. The pairs (f_i, x_i) are then plotted to yield the quantile plot in the Figure J-8.

J-5. Normal Probability Plots (Quantile-Quantile Plots). There are two types of quantile-quantile plots or *q-q* plots: an empirical quantile-quantile plot and a theoretical quantile-quantile plot. A normal probability plot is an extension of these *q-q* plots.

J-5.1. Empirical Quantile-Quantile Plot. A plot of the quantiles of two variables (e.g., the quantiles of *X* versus the quantiles of *Y*).

J-5.2. Theoretical Quantile-Quantile Plot. A plot of quantiles of a set of data against the quantiles of a specific theoretical probability distribution.

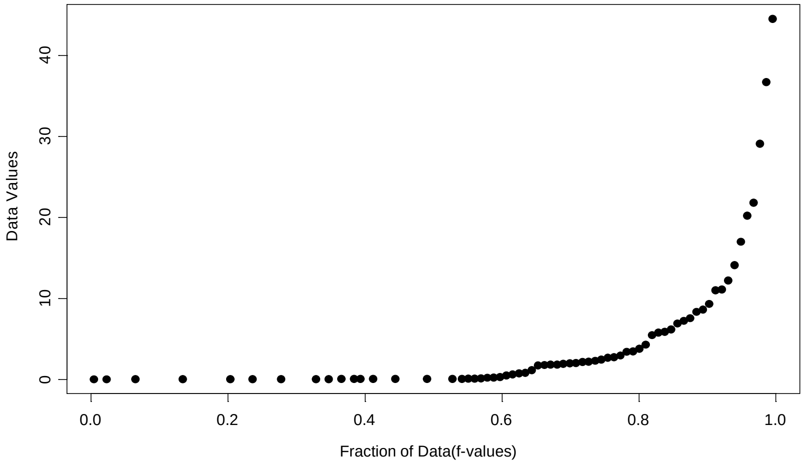


Figure J-8. Example of a Quantile Plot.

J-5.3. Normal Probability Plot. A theoretical quantile-quantile plot where the quantiles of a data set are plotted against the quantiles of the normal distribution.

J-5.4. Introduction. The following discussion will focus on the plot most commonly used for environmental data—the normal probability plot (the normal *q-q* plot); however, the discussion also 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.

J-5.4.1. A normal probability plot, as defined above, is the graph of the quantiles of a data set against the quantiles of the normal distribution (see Figure J-9). If the graph is linear, the data may be normally distributed as shown in Figure J-9a. If the graph is not linear, the departures from linearity give important information about how the data distribution deviates from a normal distribution. Further, the graph may be used to determine the degree of symmetry (or asymmetry) displayed by the data. 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 (Figure J-9b); 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 (Figure J-9c); there are more values in the tails of the data than what is expected from a normal distribution.

J-5.4.2. A normal probability plot can be used to identify potential outliers and extreme values. 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. In addition, a normal probability plot is a useful technique for identifying irregularities in the data, especially in the tails, when compared to a certain distribution.

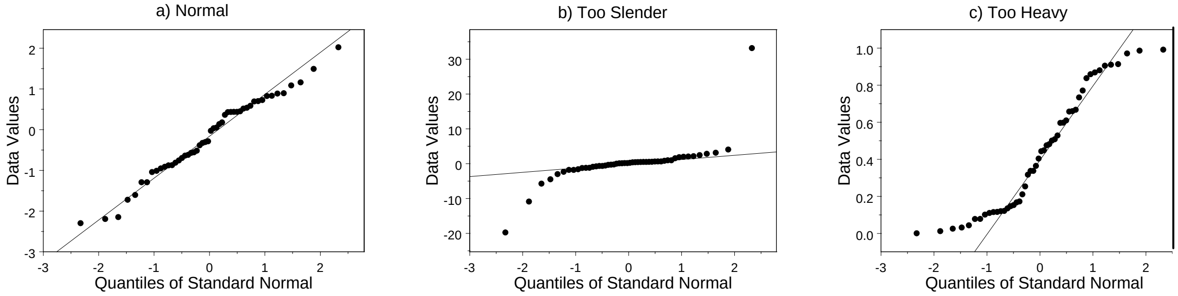


Figure J-9. Examples of Normal Probability Plots.

J-5.4.3. Directions for constructing a normal probability plot are presented in Paragraph J-5.5, followed by an example in Paragraph J-5.6.

J-5.5. Directions for Constructing a Normal Probability Plot. Let $x_{(1)}$, $x_{(2)}$,..., $x_{(n)}$ represent the *n* data points ordered from least to greatest. For each *i*, compute the fraction $f_i = (i - 0.5)/n$ and find the corresponding quantile for the standard normal distribution, Z_p, in Table B-15 of Appendix B. The normal probability plot is a plot of the pairs (Z_p, $x_{(i)}$). If the data are normally distributed, the points will fall approximately on a straight line. The slope of the line is an estimate the population standard deviation and the y-intercept (at $Z = 0$) is an estimate of the population mean, because $X = \sigma Z + \mu$.

J-5.6. Example for Constructing a Normal Probability Plot. Again, consider the following results of benzene concentrations (in µg/L) in groundwater: 0.0292, 0.0300, 0.0300, 0.0300, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0353, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0375, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0401, 0.0444, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0465, 0.0469, 0.0469, 0.0469, 0.0480, 0.0504, 0.0504, 0.0504, 0.0548, 0.0585, 0.0605, 0.0605, 0.0605, 0.0641, 0.0641, 0.0641, 0.0641, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0759, 0.0786, 0.0786, 0.0855, 0.0970, 0.0971, 0.1430, 0.2220, 0.2350, 0.3080, 0.4840, 0.6350, 0.7590, 0.8130, 1.1500, 1.7200, 1.7800, 1.8400, 1.8500, 1.9200, 2.0000, 2.0100, 2.1700, 2.1900, 2.3100, 2.4600, 2.6800, 2.7500, 2.9500, 3.4200, 3.4500, 3.7900, 4.3000, 5.4700, 5.7700, 5.8700, 6.1700, 6.9100, 7.2400, 7.5600, 8.3400, 8.6400, 9.3300, 11.0000, 11.1000, 12.2000, 14.1000, 17.0000, 20.2000, 21.8000, 29.1000, 36.7000 and 44.5000.

J-5.6.1. The data, ordered from smallest to largest, are shown below in the first column of the table ($x_{(i)}$) and the ordered number for each observation (*i*) is shown in the second column. The third column displays the values f_i for each value of *i*, where $f_i = (i - 0.5)/n$. The fourth column displays the corresponding percentiles of the standard normal distribution, Z_p ($p = f_i$).

Table J-3. Normal Probability Data

$x_{(i)}$ (Mg/L)	***i***	f_i	Z_p
0.0292	1	0.0046	–2.60
0.0300	2	0.014	–2.20
0.0300	3	0.023	–1.99
.	.	.	.
.	.	.	.
.	.	.	.
29.100	106	0.9769	1.99
36.700	107	0.9861	2.20
44.500	108	0.9954	2.61

J-5.6.2. The pairs (Z_p, $x_{(i)}$) are then plotted to yield the normal probability plot shown in Figure J-10. Because this plot is clearly nonlinear, these data are unlikely to be from a normal distribution

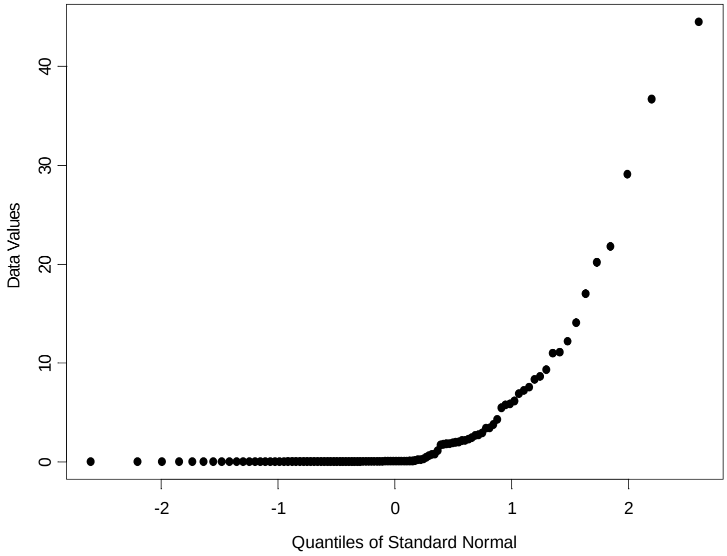


Figure J-10. Example of a Normal Probability Plot.

J-6. Empirical Quantile-Quantile Plots. 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, plotting the distribution of two variables instead of the distribution of one variable in relation to a fixed distribution.

J-6.1. Introduction. If the distributions are roughly the same, the graph will be approximately linear; the slope will be nearly one and the intercept will be nearly zero. If the distributions are not the same, then the graph will not necessarily be 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 is 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. Directions for constructing an empirical *q-q* plot are presented in Paragraph J-6.2 followed by an example in Paragraph J-6.3.

J-6.2. Directions for Constructing an Empirical q-q Plot. Let $x_1, x_2, \ldots, x_n$ represent n data points of one variable and let $y_1, y_2, \ldots, y_m$ represent a second variable of m data points.

J-6.2.1. Let $x_{(i)}$, for $i = 1$ to n, be the first sample listed in order from smallest to largest so that:

$x_{(1)}$ ($i = 1$) is the smallest

$x_{(2)}$ ($i = 2$) is the second smallest

$x_{(n)}$ ($i = n$) is the largest.

J-6.2.2 Let $y_{(i)}$, for $i = 1$ to m, be the second sample listed in order from smallest to largest so that:

$y_{(1)}$ ($i = 1$) is the smallest

$y_{(2)}$ ($i = 2$) is the second smallest

$y_{(m)}$ ($i = m$) is the largest.

J-6.2.3. 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. Because $n = m$, replace m by n. A plot of the following pairs is an empirical *q-q* plot:

$(x_{(1)}, y_{(1)}), (x_{(2)}, y_{(2)}), \ldots, (x_{(n)}, y_{(n)})$.

J-6.2.4. If the two variables have a different number of observations ($n > m$), then the empirical *q-q* 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 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.

J-6.2.5. If *g* = 0, plot the pair ($y_{(i)}, x_{(i)}$). Otherwise, plot the pair $\left(y_i, \left(1 - g\right)x_j + g\, x_{(j+1)}\right)$. A plot of these pairs is an empirical *q-q* plot.

J-6.3. Example for Constructing an Empirical q-q Plot. Consider the following arsenic concentrations in subsurface soil samples (mg/kg): 2.15, 2.26, 2.37, 2.18, 1.93, 2.06, 2.00, 1.42, 1.31, 1.95, 2.88, 1.71, 1.92, 2.33, 1.55, 1.75, 2.09, 2.38, 2.11, 2.33, 1.98, 1.55, 1.76, 1.31, 2.34, 1.22, 1.81, 1.91, 2.31, 2.10, 1.89, 1.91, 1.49, 1.79, 2.71, 1.70, 1.93, 1.64, 1.94, 3.15, 2.32, 1.31, 1.97 and 1.48. And the following chromium concentrations in subsurface soil samples (mg/kg) are: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, 3.84, 2.95, 5.17, 4.80, 4.53, 4.01, 5.91, 3.96, 4.81, 5.27, 5.99, 4.60, 5.51, 4.72, 3.56, 4.22, 3.91, 5.81, 4.48, 5.10 ,4.94, 4.76, 4.62, 4.72, 4.73, 3.21, 4.14, 4.85, 4.25, 5.09, 3.68, 5.12, 6.60, 6.19, 3.15, 4.11 and 2.80.

J-6.3.1. An empirical *q-q* plot will be used to compare the distributions of these two analytes. As there are 44 observations of arsenic and 44 observations of chromium, the case for *m* = *n* will be used. Therefore, for *i* = 1, 2, ..., 44, compute:

$$(x_{(1)}, y_{(1)}), (x_{(2)}, y_{(2)}), \ldots (x_{(44)}, y_{(44)}) .$$

J-6.3.2. These pairs are plotted below, along with the best fitting regression line, as shown in Figure J-11.

J-7. Plots for Temporal Data.

J-7.1. Introduction. Data collected over specific time intervals (such as monthly, bi-weekly, 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, or cyclical patterns, or all three. Some graphical representations specific to temporal data are the time plot, correlogram, and variogram.

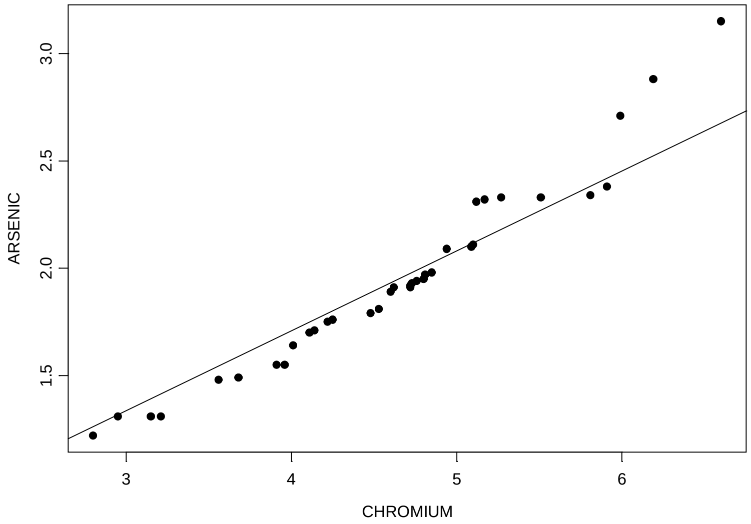


Figure J-11. Empirical *q-q* Plot.

J-7.1.1. Time Plot. This is a plot of time versus some variable (e.g., concentration).

J-7.1.2. Time Series Plot. This is a time plot in which measurements of a variable are taken at regular, fixed intervals over time.

J-7.1.3. Correlogram. This is a plot that displays serial correlation when the data are collected at equally spaced time (or distance) intervals.

J-7.1.4. Variogram. This is a plot that displays the same information as a correlogram except that the data may be based on unequally spaced time (or distance) intervals. Further discussion of the variogram is contained in Appendix R.

J-7.2. Discussion. Data collected at regular time intervals are called time series. The graphical representations presented in this Paragraph 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 or trend analyses such as those described in Appendix Q, the graphical representations presented below will play an important role in this analysis. If the analyst decides not to use time series methodologies, these representations will help identify temporal patterns that need to be accounted for in the analysis of the data.

J-7.2.1. The analyst examining temporal environmental data may be interested in cyclic trends, directional trends, serial correlation, and stationarity.

J-7.2.1.1. Cyclic Trend. This is a pattern in the data (e.g., attributable to seasonal changes) that repeats over time.

J-7.2.1.2. Directional Trend. This is a downward or upward trend in the data.

J-7.2.1.3. Serial Correlation. This is a measure of the extent to which successive observations are related.

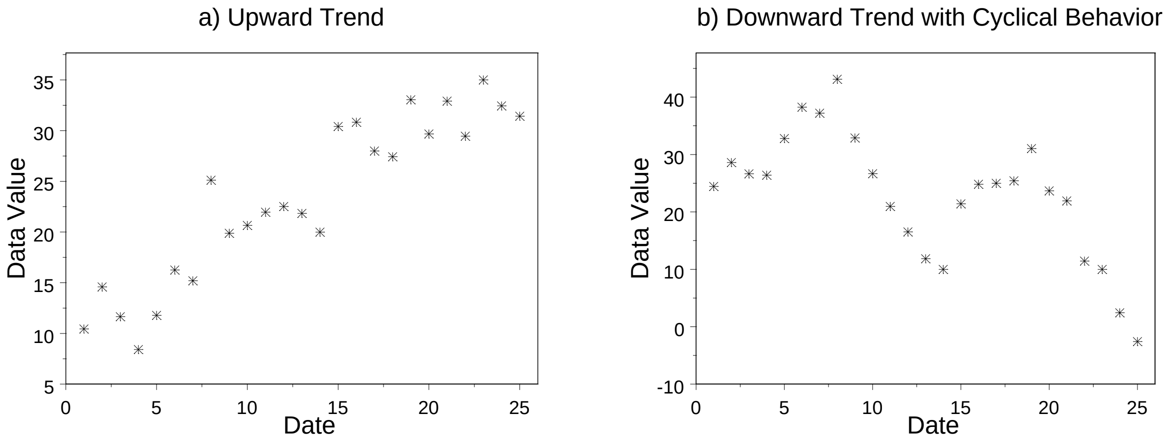
J-7.2.1.4. Stationarity. This describes the situation when the data looks the same over all time periods.

J-7.2.2. Cyclic trends repeat over time; the data rise and fall regularly over one or more time periods. These 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, 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 they are related, statistical quantities calculated without accounting for it may be biased.

J-7.2.3. Another issue for temporal data is stationarity. Stationary data look the same over all time periods. Directional or cyclical trends and increasing (or decreasing) variability among the data imply that the data are not stationary. Temporal data are sometimes used in environmental projects along with a statistical hypothesis test to determine if contaminant levels have changed over time. If the hypothesis test does not account for temporal trends or seasonal variations, the data must achieve a steady state before the hypothesis may be tested. The data must be essentially the same for comparable periods of time both before and after the hypothesized time of change.

J-7.2.4. Sometimes multiple observations are taken in each time period. For example, the sampling design may specify selecting five samples every Monday for 3 months. If this is the case, the time plot 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 changing between time periods. In addition, each time period may be treated as a distinct variable and the methods for plots for two or more variables may be applied.

J-7.3. Time Plots. One of the simplest plots to generate that provides a large amount of information is a time plot. This is a plot of the data that makes it easy to identify large- and small-scale trends over time. Small-scale trends show up on a time plot as fluctuations in smaller (or shorter) 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 show directional trends and increased variability over time. Possible outliers may also be easily identified using a time plot. Figure J-12 displays two examples of time plots. Figure J-12a demonstrates an upward trend, while Figure J-12b shows a downward trend superimposed with cyclical behavior.


a) Upward Trend

b) Downward Trend with Cyclical Behavior

Figure J-12. Examples of Time Plots.

J-7.3.1. Discussion. A time plot 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. Although the points plotted on a time plot may be joined by lines, 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 Paragraph J-7.3.2 along with an example.

J-7.3.2. Directions for Generating a Time Plot. 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.

J-7.3.2.1. Consider the following 15 benzene concentrations (μg/L) measured in groundwater (listed in order by day): 12.200, 3.790, 3.420, 5.470, 0.813, 1.840, 7.560, 4.300, 2.680, 6.170, 0.635, 2.190, 1.720, 1.150 and 0.484.

J-7.3.2.2. 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 = the number of the day, and x_i = the concentration level.

J-7.3.2.3. A time plot of these data is shown in Figure J-13.

J-7.4. Plot of the Autocorrelation Function (Correlogram).

J-7.4.1. Discussion. 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 first autocorrelation coefficient (r_1) is the correlation between all points that are one time unit (k_1) apart; the second autocorrelation coefficient (r_2) is the correlation between points that are two time units (k_2) apart; and so on. A correlogram (Figure J-14) is a plot of the sample autocorrelation coefficients in which the values of k versus the values of r_k are displayed.

J-7.4.1.1. The correlogram is used for modeling time series data and helps to determine if serial correlation is large enough to create problems in the analysis of temporal data using other methodologies. A quick method for determining if serial correlation is large is to place horizontal lines at $\pm 2/n$, where n is the number of samples on the correlogram (shown as horizontal lines on Figure J-14). Autocorrelation coefficients that exceed this value require further investigation.

J-7.4.1.2. 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 (or distance) intervals. For more information on the construction and uses of the variogram, consult a statistician.

J-7.4.1.3. Directions for constructing a correlogram are contained in Paragraph J-7.4.2, followed by example calculations in Paragraph J-7.4.3.

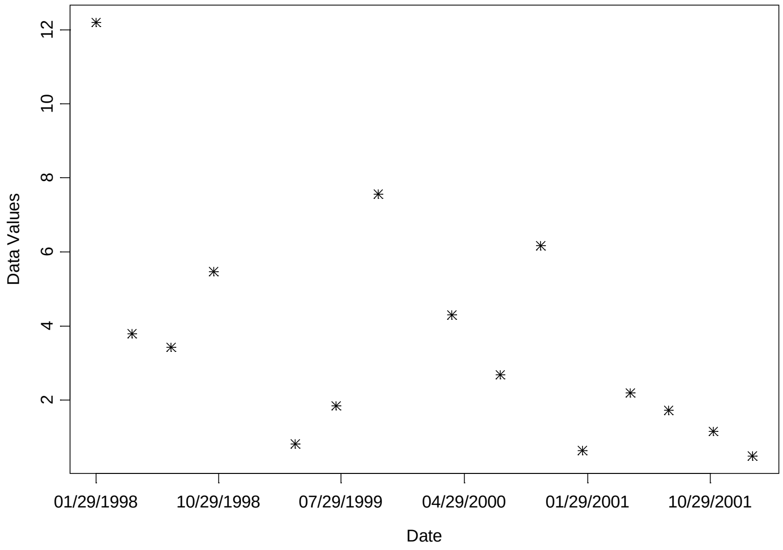
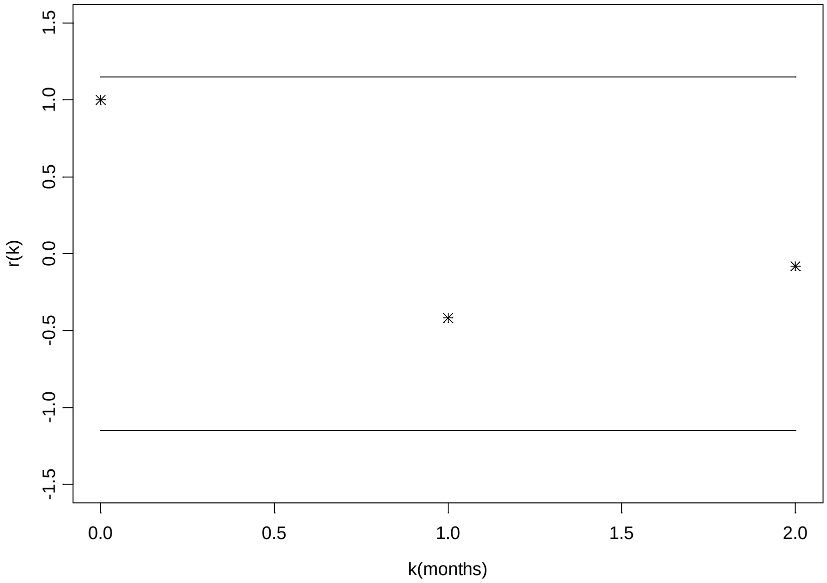


Figure J-13. Example Time Plot.



- Figure J-14. Correlogram for Data in Paragraph J-7.4.3.

J-7.4.2. 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.

J-7.4.2.1. To construct a correlogram, first compute the sample autocorrelation coefficients. So for k = 0,1, ….., compute r_k where

$$r_k = \frac{g_k}{g_o}$$

and

$$g_k = \frac{\sum_{t=k+1}^{n}(x_t - \overline{x})(x_{t-k} - \overline{x})}{n} \ .$$

J-7.4.2.2. Once r_k has been computed, a correlogram is the graph (r_k, k) for *k* = 0, 1, …

J-7.4.2.3. Compute up to approximately *k* = *n*/6.

J-7.4.2.4. Also, note that $r_0 = 1$.

J-7.4.2.5. Finally, place horizontal lines at

$$\pm \frac{2}{\sqrt{n}}.$$

J-7.4.3. Example for Constructing a Correlogram. A correlogram will be constructed using the following three benzene concentrations in groundwater, collected monthly—month 1: 11.10 ppb, month 2: 2.46 ppb, month 3: 5.77 ppb. Although a correlogram would not typically be constructed when only three data points are available, only three data points are used here so that all computations may be shown. The rules that up to *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 (Appendix D), which is 6.44 for the three points. Then,

$$g_0 = \sum_{t=1}^{3}(x_t - \overline{x})(x_{t-0} - \overline{x}) = \frac{\sum_{t=1}^{3}(x_t - \overline{x})^2}{3} = \frac{(11.10-6.44)^2 + (2.46-6.44)^2 + (5.77-6.44)^2}{3}$$

$$= \frac{21.72 + 15.84 + 0.45}{3} = 12.67$$

$$g_1 = \frac{\sum_{t=2}^{3}(x_t - 6.44)(x_{t-1} - 6.44)}{3} = \frac{(x_2 - 6.44)(x_1 - 6.44) + (x_3 - 6.44)(x_2 - 6.44)}{3}$$

$$= \frac{(2.46 - 6.44)(11.10 - 6.44) + (5.77 - 6.44)(2.46 - 6.44)}{3} = -5.29$$

$$g_2 = \frac{\sum_{t=3}^{3}(x_t - 6.44)(x_{t-2} - 6.44)}{3} = \frac{(x_3 - 6.44)(x_1 - 6.44)}{3} = \frac{(5.77 - 6.44)(11.10 - 6.44)}{3} = -1.05 \ .$$

So,

$$r_1 = \frac{g_1}{g_0} = \frac{-5.29}{12.67} = -0.418$$

$$r_2 = \frac{g_2}{g_0} = \frac{-1.04}{12.67} = -0.082 .$$

Remember, r_0 = 1. Thus, the correlogram of these data is a plot of (0, 1) (1, –0.418) and (2, –0.082) with two horizontal lines at (±1.15). This graph is shown in Figure J-14. In this case, it appears that the observations are not serially correlated because all of the correlogram points are within the bounds of (±1.15).

J-7.4.4. Multiple Observations Per Time Period. In environmental data collection, multiple observations are sometimes taken for each time period. For example, the data collection design may specify collecting and analyzing five samples from a drinking well every Wednesday for 3 months. If this is the case, a time plot 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 changes. A time plot of the means will allow the analyst to determine if they are changing between the collection periods. In addition, each collection period may be treated as a distinct variable and the methods applied as described in the section on plots for two or more variables (Paragraph J-9).

J-8. Plots for Spatial Data.

J-8.1. Introduction. The graphical representations of the preceding Paragraphs may also be useful for exploring spatial data. An analyst examining spatial data may be interested in locating extreme values, overall spatial trends, and the degree of continuity among neighboring locations. Graphical representations for spatial data include postings, symbol plots, and correlograms (the correlograms would be generated by collecting samples at equally spaced sampling locations). The graphical representations presented below are recommended for all spatial data regardless of whether or not geostatistical methods will be used to analyze it. They will help identify spatial patterns that need to be accounted for in the analysis of the data. If geostatistical methods such as kriging are used to analyze the data, these methods will play an important role.

J-8.2. Posting Plots. A posting plot (Figure J-15) 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 Paragraph J-8.4.

J-8.3. Symbol Plots. For large amounts of data, a posting plot may not be feasible and a symbol plot (Figure J-16) may be used. A symbol plot is 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 Paragraph J-8.4.

J-8.4. Directions for Generating a Posting Plot and Symbol Plot with an Example.

J-8.4.1. Directions. 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. The Posting plot and the Symbol plot are displayed as Figures J-15 and J-16.

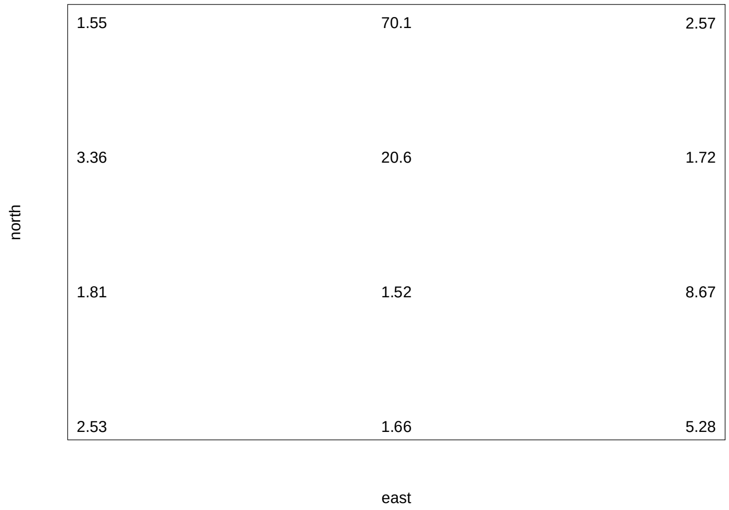


Figure J-15. Posting Plot.

J-8.4.2. Example. The spatial data displayed in Table J-4 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.

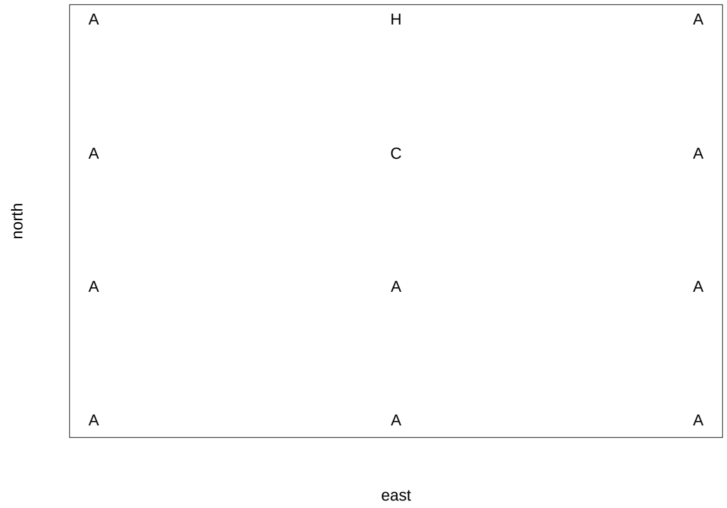


Figure J-16. Symbol Plot.

Table J-4. Spatial Data

Range	Symbol	Range	Symbol
0.0–9.9	A	40.0–49.9	E
9.9–19.9	B	50.0–59.9	F
20.0–29.9	C	60.0–69.9	G
30.0–39.9	D	70.0–79.9	H

Northing	Easting	*C*	Symbol	Northing	Easting	*C*	Symbol
25.0	0.0	2.53	A	15.0	15.0	2.57	A
25.0	5.0	1.81	A				
25.0	10.0	3.36	A				
25.0	15.0	1.55	A				
20.0	0.0	1.66	A				
20.0	5.0	1.52	A				
20.0	10.0	20.60	C				
20.0	15.0	70.10	H				
15.0	0.0	5.28	A				
15.0	5.0	8.67	A				
15.0	10.0	1.72	A				

J-8.5. Other Spatial Graphical Representations. The two plots discussed above, posting and symbol, 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 mostly used for preliminary geostatistical analysis. These graphs can be difficult to develop and interpret. For more information on these, consult a statistician.

J-8.5.1. An *h* scatter plot 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, an *h* scatter plot could be based on all the pairs whose locations are 1 meter apart in a southerly direction. An *h* scatter plot is similar in appearance to a scatter plot. The shape of the spread of the data in an *h* scatter plot indicates the degree of continuity among data values a certain distance apart in a 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.

J-8.5.2. A correlogram is a plot of the correlations of the *h* scatter plots. Because the 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 graph of how these correlations change for different separation distances in a fixed direction. The correlogram is such a plot and allows the analyst to evaluate the change in correlation in a fixed direction as a function of the distance between two points. A spatial correlogram is similar in appearance to a temporal one. It 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. Correlograms for spatial data are related to the semivariograms discussed in Appendix R.

J-8.5.3. 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 be 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 judgment to adequately interpret the findings.

J-9. Visualizing Higher Dimensional Data: Plots for Two or More Variables.

J-9.1. Introduction. To compare and contrast several variables, collections of the single variable displays described previously are useful. For example, the analyst may generate side-by-side box-and-whiskers plots or histograms for each variable using the same axis for all of the variables.

J-9.1.1. Figure J-17 illustrates side-by-side box-and-whiskers plots for naphthalene concentrations at various groundwater-monitoring wells at a given site.

J-9.1.2. In addition, the number of detected observations over the total number of observations has been placed towards the top of the graph. 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 histograms back to back. In addition, some special plots have been developed to display two or more variables; these allow comparison and contrast of individual data points of all the variables. These plots are described below.

J-9.2. Plots for Individual Data Points.

J-9.2.1. As it is difficult to visualize data in more than two or three 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. Such plots include Profiles, Glyphs, and Stars (Figure J-18). 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 zero, the maximum be of length one. The remaining values of each variable are then proportioned, based on the magnitude of each value in relation to the minimum and maximum.

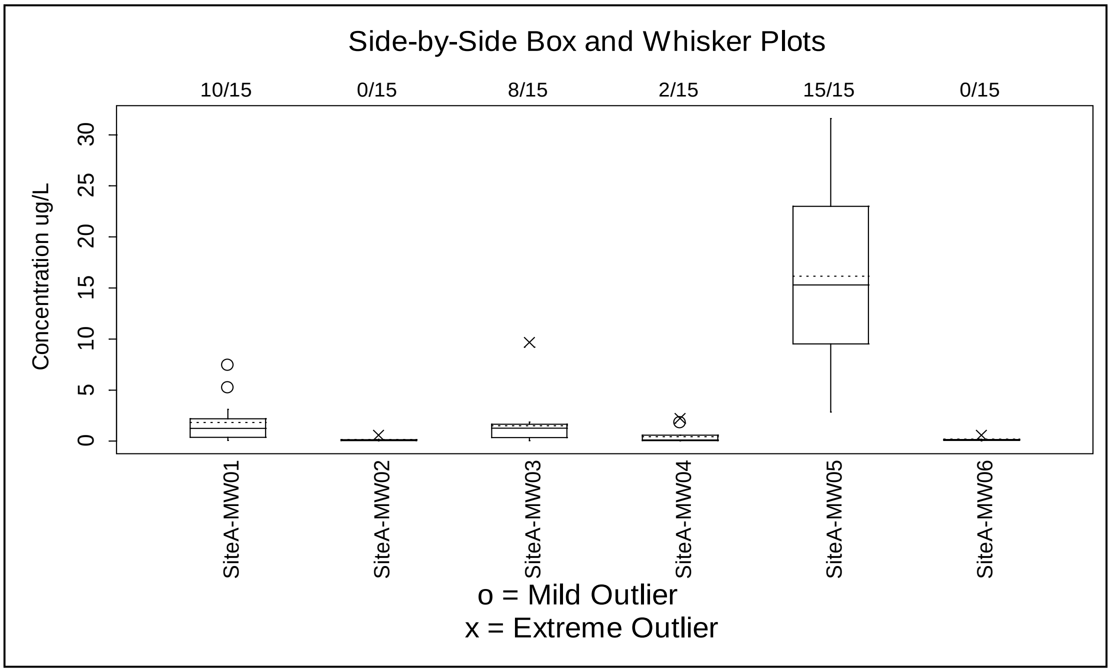


Figure J-17. Concentrations of Naphthalene at Site A Wells.

J-9.2.2. 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.

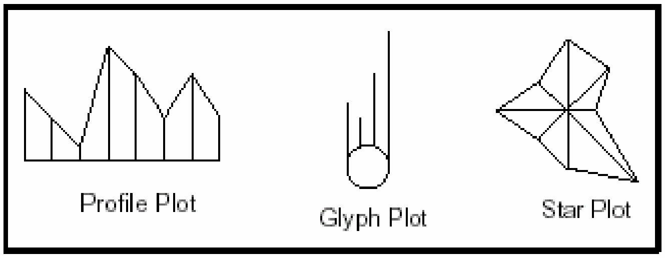


Figure J-18. Graphical Representations of Multiple Variables.

J-9.3. 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 J-19) and many computer graphics packages can construct three-dimensional scatter plots. Directions for constructing a scatter plot for two variables are given in Paragraph J-9.4 along with an example in Paragraph J-9.5.

J-9.3.1. 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 is approximately equal to the square of the other. A scatter plot of these 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.

J-9.3.2. Additional information can be placed in a scatter plot. Labels can be placed on each value in the scatter plot to identify the sample location of a value. Different colors or symbols may be used to identify unique groupings of the data. For example, the scatter plot data may contain concentrations from multiple sampling events, with a unique symbol used to identify each event. This will show trends in concentrations as well as distinguishing sampling events.

J-9.4. Directions for Generating a Scatter Plot. Let x_1, x_2,..., x_n represent one variable of n data points and let y_1, y_2,..., y_n represent a second variable of the same 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.

J-9.5. Example of a Scatter Plot. A scatter plot is prepared for arsenic and chromium concentrations in subsurface soil at Site A, using the data in Table J-5. Arsenic values are shown on the horizontal axis and chromium values are displayed on the vertical axis of Figure J-19.

J-9.6. Extensions of the Scatter Plot. It is easy to construct a two-dimensional scatter plot manually. Many software packages can construct useful two- and three-dimensional scatter plots. However, it is difficult to construct and interpret a scatter plot for more than three variables, so several graphical representations have been developed that extend the idea of a scatter plot to data consisting of two or more variables.

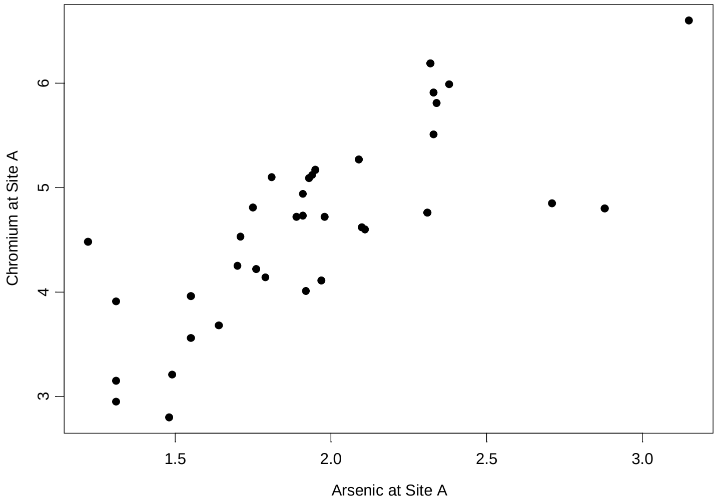


Figure J-19. Example of a Scatter Plot.

J-9.7. Scatter Plot Matrix. A scatter plot matrix is a useful method for extending scatter plots to higher dimensions. In this case, a scatter plot is developed for all possible pairs of the variables that are then displayed in a matrix format. This method is easy to use and is a concise method of displaying the individual scatter plots. However, this method does not contain information on three-way or higher interactions between variables. An example of a scatter plot matrix is contained in Figure J-20.

J-9.8. Side-by-Side Scatter Plot. A form of scatter plot, called a side-by-side scatter plot, is designed in a manner similar to the side-by-side box-and-whiskers plots presented earlier. Such scatter plots are developed using the horizontal axis as a label for each variable and using the vertical axis as the range of values for the variables. Figure J-21 illustrates a side by side scatter plots for the same data presented in Figure J-5. In Figure J-21, the *y*-axis is the range of concentrations for naphthalene and the *x*-axis represents the wells that were sampled during the site investigation. Because the wells were sampled over several years, different symbols are used to represent each year—triangles represent 1998, squares represent 1999, and circles represent 2000. In addition, because there are detected and non-detected results in the data, open symbols were used for non-detected values and closed symbols were used for detected values. At the top of the graph, a ratio is shown that states the number of detected observations over the total number of observations for each well sampled. A side-by-side scatter plot can be a useful tool in comparing and contrasting concentrations of a specific chemical at various data points (e.g., different wells at a particular site).

Table J-5.
Arsenic and Chromium Concentrations in Subsurface Soil at Site A

Sample ID	Arsenic (mg/kg)	Chromium (mg/kg)	Sample ID	Arsenic (mg/kg)	Chromium (mg/kg)
APA-EPC-SB01-030	1.31	2.95	APA-EPC-SB07-030	1.81	5.1
APA-EPC-SB01-040	1.95	5.17	APA-EPC-SB07-040	1.91	4.94
APA-EPC-SB01-050	2.88	4.8	APA-EPC-SB07-050	2.31	4.76
APA-EPC-SB02-030	1.71	4.53	APA-EPC-SB08-030	2.1	4.62
APA-EPC-SB02-040	1.92	4.01	APA-EPC-SB08-040	1.89	4.72
APA-EPC-SB02-050	2.33	5.91	APA-EPC-SB08-050	1.91	4.73
APA-EPC-SB03-030	1.55	3.96	APA-EPC-SB09-030	1.49	3.21
APA-EPC-SB03-040	1.75	4.81	APA-EPC-SB09-040	1.79	4.14
APA-EPC-SB03-050	2.09	5.27	APA-EPC-SB09-050	2.71	4.85
APA-EPC-SB04-030	2.38	5.99	APA-EPC-SB10-030	1.7	4.25
APA-EPC-SB04-040	2.11	4.6	APA-EPC-SB10-040	1.93	5.09
APA-EPC-SB04-050	2.33	5.51	APA-EPC-SB10-050	1.64	3.68
APA-EPC-SB05-030	1.98	4.72	APA-EPC-SB11-030	1.94	5.12
APA-EPC-SB05-040	1.55	3.56	APA-EPC-SB11-040	3.15	6.6
APA-EPC-SB05-050	1.76	4.22	APA-EPC-SB11-050	2.32	6.19
APA-EPC-SB06-030	1.31	3.91	APA-EPC-SB12-030	1.31	3.15
APA-EPC-SB06-040	2.34	5.81	APA-EPC-SB12-040	1.97	4.11
APA-EPC-SB06-050	1.22	4.48	APA-EPC-SB12-050	1.48	2.8

J-9.9. Parallel Coordinate 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 set of *m*-dimensional sample points $x_i = (x_{1i}, x_{2i}, x_{3i},..., x_{mi})$, where $i = 1, 2, 3…n$. For the i^{th} *m*-dimensional point, the variable $X_1 = x_{1i}$, $X_2 = x_{2i}$ and so forth. A parallel coordinate plot is constructed by first placing an axis (X_i) for each of the *m* variables parallel to each other. Each point x_{i} is graphically represented by plotting x_{1i} on the X_1 axis, x_{2i} on the X_2 axis and so forth, and then joining the set of *m* plotted values with a broken line. This method contains all of the information available on a scatter plot in addition to information on three-way and higher interactions (i.e., clustering among three variables). However, for *m* variables one must construct $m(m – 1)/2$ parallel coordinate plots in order to display all possible pairs of variables. For an example of a parallel coordinate plot see EPA/240/B-026/003, QA/G-9S.

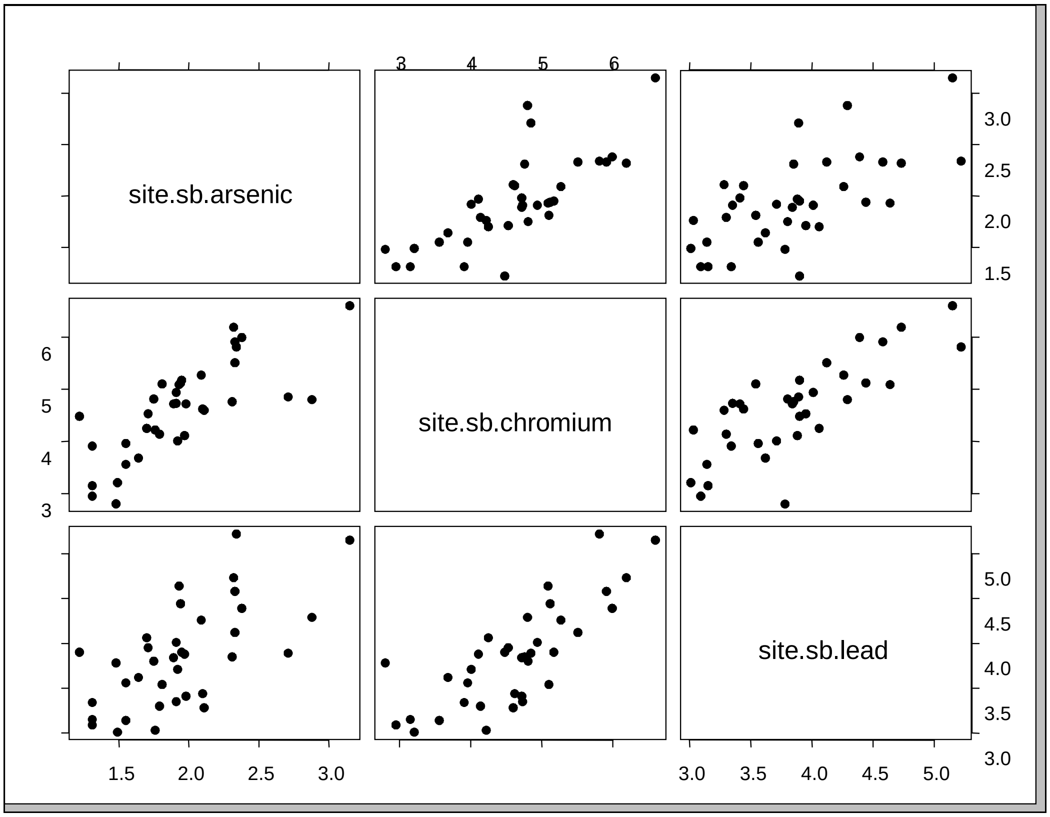


Figure J-20. Scatter Plot Matrix.

J-10. Contouring Data. Contouring site data helps in visualizing site conditions and presenting results. The results could be groundwater elevations and flow directions or locations and volumes of contamination. Contaminant concentrations are typically plotted by contouring the data over a site map. Contours or isopleths are lines of equal value (e.g., concentration). Lines or areas can be color coded or defined by a concentration range rather than a single value. Contour lines may not cross each other although they may form loops. The spacing of contour lines represents the gradient of the variable.

J-10.1. A topographic elevation map is a common contour map. Environmental data such as water table drawdown or chemical concentrations in water and air readily lend themselves to contouring. Contour maps are useful in data analysis because changes over distance, gradients, hot-spots, and the location of contaminants relative to site features, such as buildings and site boundaries, are apparent.

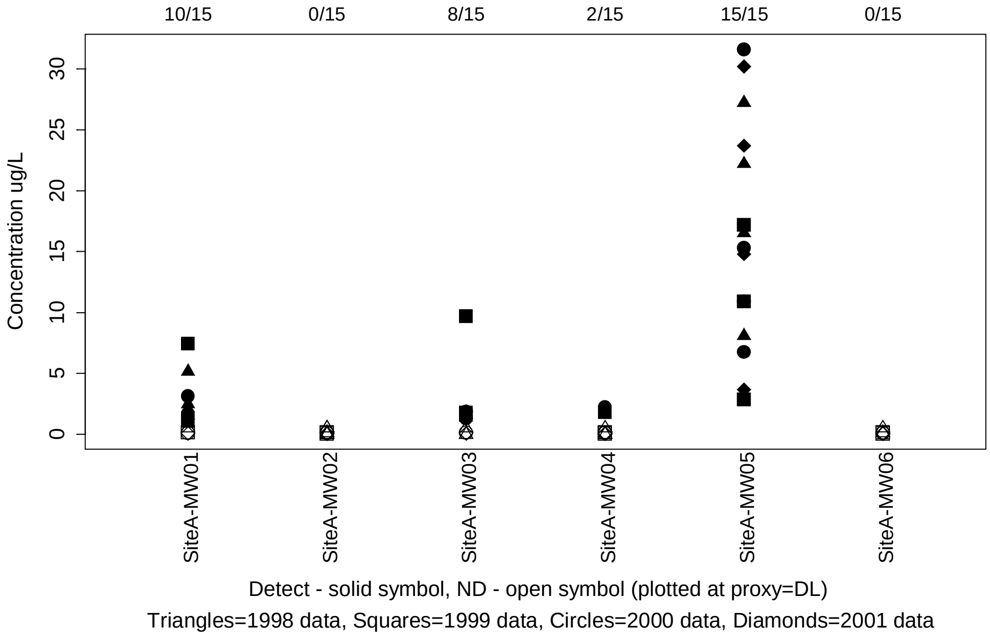


Figure J-21. Naphthalene Concentrations at Site A Wells: Side-by-side Scatter Plots.

J-10.2. Contours must be interpolated because data coverage at a site is partial. For instance, water levels are measured only in monitoring wells even though the water table exists between wells. Interpolation estimates values within the existing data set while extrapolation estimates values outside the existing data set. Objects or values that are discontinuous spatially do not lend themselves to interpolation, for example, the presence of unexploded munitions at a test range.

J-10.3. There are numerous interpolation techniques available and selecting which to use depends on the media (soil, water, or air) and site-specific circumstances. Interpolation methods must be evaluated for their applicability, artifacts, and accuracy based on the analyst's site knowledge and technical expertise. Small data sets and software default values can result in contour maps that do not reflect actual site conditions. Software will often attempt to extrapolate beyond the data coverage unless a boundary is established or settings are carefully selected. It is good practice to contour data by hand and then compare results to computer-generated output. This allows the analyst to incorporate site-specific knowledge and intuition.

J-10.4. Currently available contouring software facilitates data interpretation and reinterpretation. Because data may be stored electronically, they may be readily revised and sorted. Numerous interpolation methods can be experimented with quickly. Pertinent information (such as sample depth, soil type, concentration) for a sampling point can be viewed by placing the cursor over it. High-end two-dimensional (2-D) and three-dimensional (3-D) color graphic images plus animation can be generated.

J-10.5. 3-D contouring, which is typically done with the aid of computers, is important to consider as an analysis tool. 3-D iso-surfaces are generated in lieu of 2-D contour lines. The failure to view contamination in natural systems in three dimensions, excluding the vertical or depth components, can adversely affect decision-making. It can give rise to misinterpretations of contamination sources (responsible parties) and transport, particularly when the geology is not laterally homogeneous or contaminants have densities different from the transport medium (solvents denser than water).

J-10.6. For field data to be adequately characterized, the manner in which the data will be analyzed should be considered when designing the sampling plan. The study area or area of concern should be well within the sample grid. This helps establish a boundary and ensure that measurements will be taken where they are most needed. The spacing of sampling locations affects the manner in which data will be analyzed and what can be learned from the data set. Poorly sized sampling grids can miss hot-spots or make the site appear more contaminated than it actually is. Poorly distributed data will lead to software drawing concentric contours around known values. It is often the case that vertical sample spacing is closer than the horizontal spacing. This situation can cause the vertical samples to unduly override the horizontal characteristics of the subsurface. Scaling features can be used to compensate for biased data sets.

J-10.7. No single interpolation method will be universally appropriate. In addition to trying more than one interpolation method, it is advisable to examine the computation used by the software. Some methods are better suited for certain data sets, such as those where values go from one extreme to another quickly or those where the changes are gradual and smooth. The mathematical function can also limit the interpolated value to a value not necessarily representative of site conditions. For example, simple inverse distance weighting (IDW) interpolates using the mean of two known values. The result is that the interpolated value lies between both known values and minimum and maximum values are not derived. It is also possible to interpolate negative values. Understanding the mathematical functions allows the input variables to be adjusted for individual circumstances. For instance, truncating a data set by setting a minimum and maximum concentration can alleviate some problems. The weight an interpolated point receives is directly related to its proximity to a known point. An interpolation’s accuracy can be checked by randomly removing data points and then comparing the new interpolated value to the value that was removed.

J-10.8. Interpolation methods include the following.

J-10.8.1. Linear Interpolation. This is the mathematically simplest interpolation technique. This technique is referred to as manual or hand interpolation or contouring. A straight line drawn between two known values is subdivided into equal segments. The location of the estimated value is calculated using proportions.

J-10.8.2. IDW Interpolation. This gives more weight to an estimated point the closer it is to a known data point. IDW is often used for groundwater level data. A power value of two typically yields smooth contours.

J-10.8.3. “Natural Neighbor” Interpolation. This uses the same mathematical equation as IDW but the weighting technique is different. In addition, a polygon network is employed rather than a triangle network. The natural neighbor method may work well with clustered data.

J-10.8.4. Triangular Irregular Network (TIN) Interpolation. This connects the data points with a gridwork of triangles. TIN is used with linear interpolation to estimate values from the three vertices of each triangle.

J-10.8.5. Spline Method. This uses a polynomial function to fit a curve through the known points. It works well for data that change gradually. Splining is often applied to dense, regularly spaced data.

J-10.8.6. Kriging. This uses spatial variance to interpolate data. Kriging assumes, as IDW does, that distance and weight are related, but it also accounts for the spatial variance (spread) as a function of distance. Variograms, used in kriging, are graphs of a mathematical function that show spatial dependence in relation to distance and direction. Kriging has an intermediate step of matching the experimental variogram curve to a model variogram. Kriging handles steep gradients well, and is a good place to start for analyzing geological data because it was originally developed to predict ore locations for the mining industry. Variograms can provide insight into data sets even when kriging is not being performed.

J-10.9. Figure J-22 shows a groundwater elevation contour plot drawn by linear interpolation. Figure J-23 shows the same groundwater elevation data, factoring in the analyst’s site knowledge. Figure J-24 shows the groundwater elevation data plot drawn by modeling software using IDW.

J-10.10. By using contouring with groundwater modeling software, otherwise static contour maps can be run forward or backward in time. This predictive modeling can be used to estimate the date at which some historical contaminant was released or plume migration at some future time. Groundwater modeling, GIS, statistics, and mapping software can perform various interpolation methods.

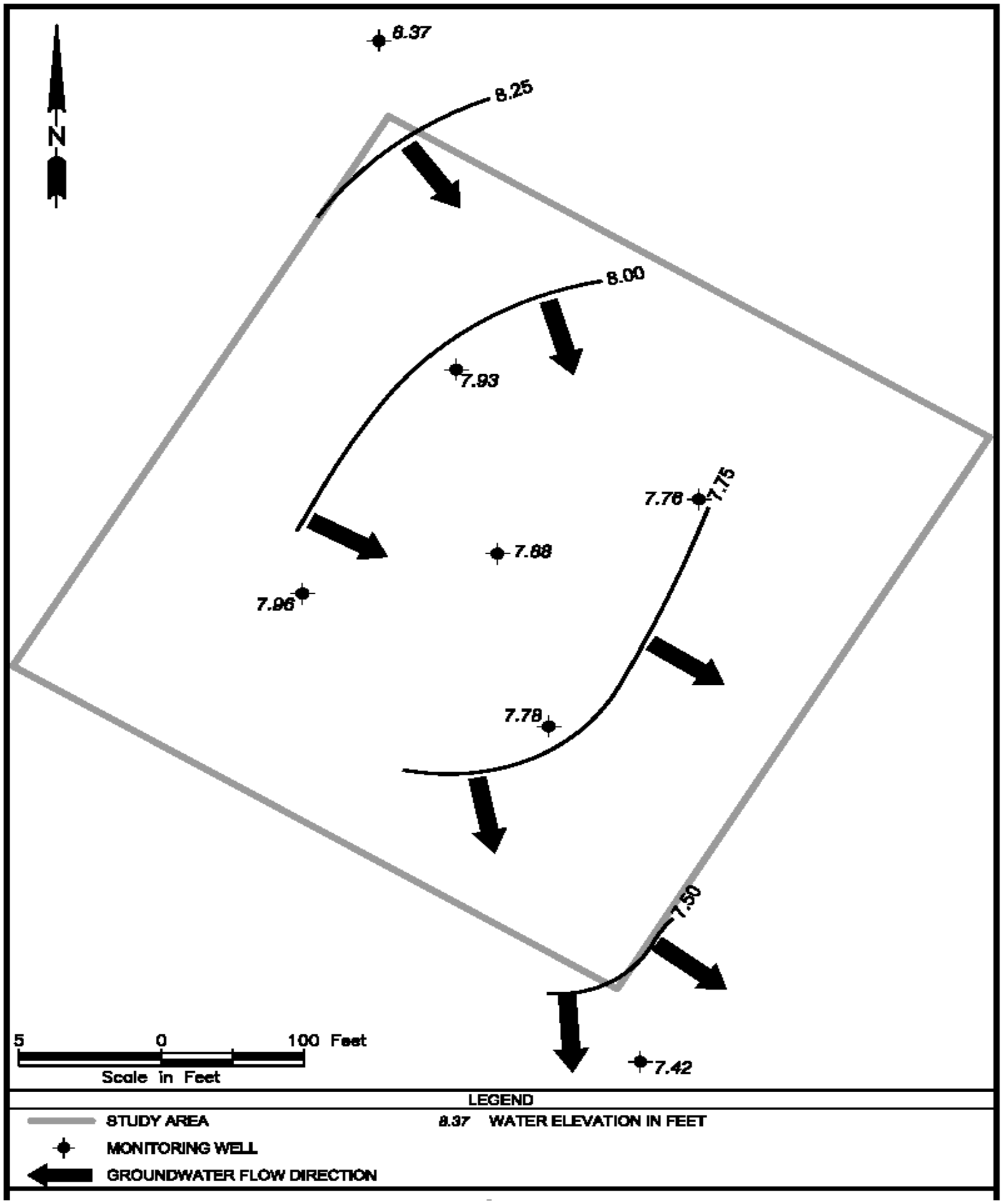


Figure J-22. Contour Plot Drawn by Linear Interpolation.

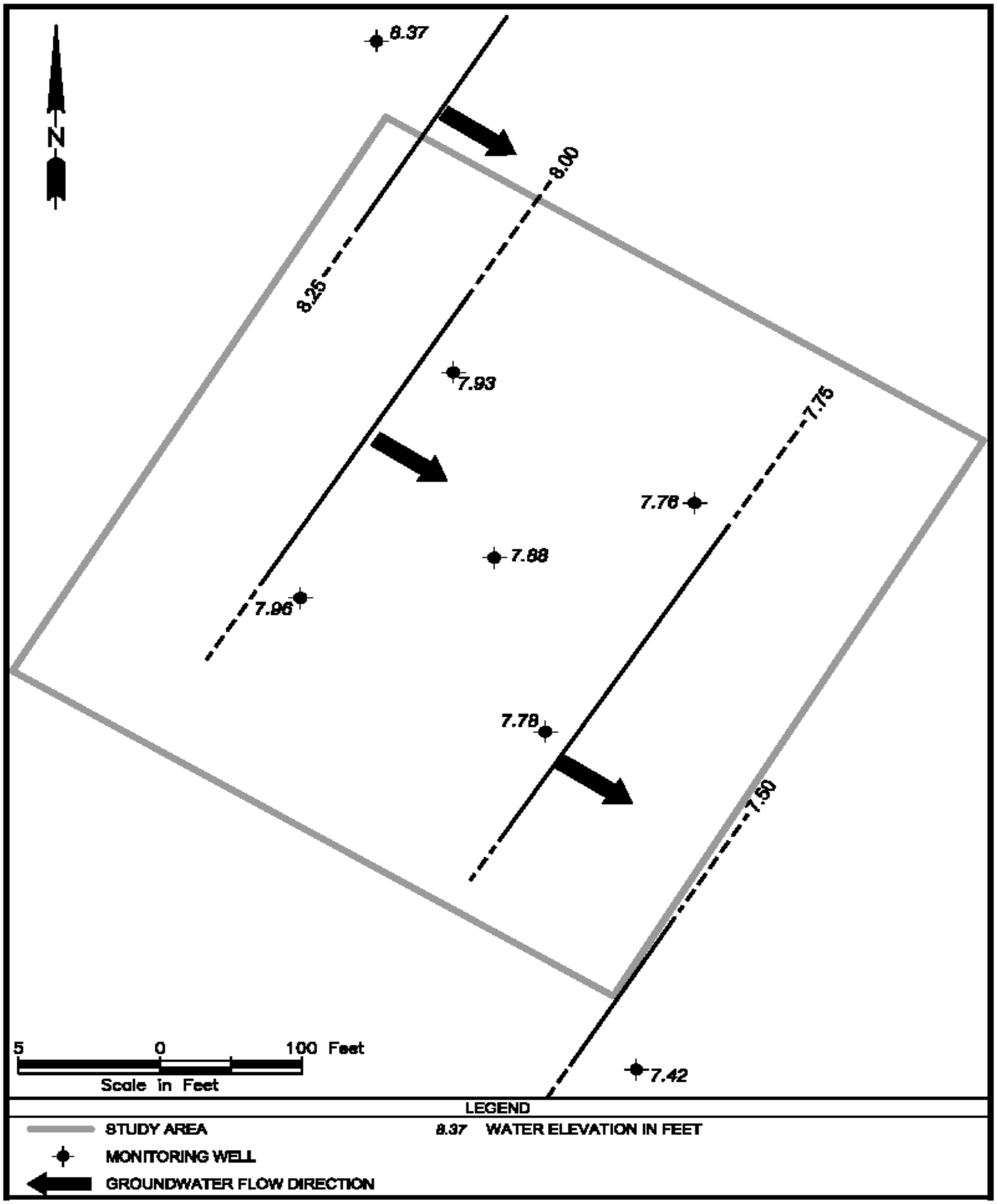


Figure J-23. Contour Plot Drawn by Linear Interpretation Incorporating Site Knowledge.

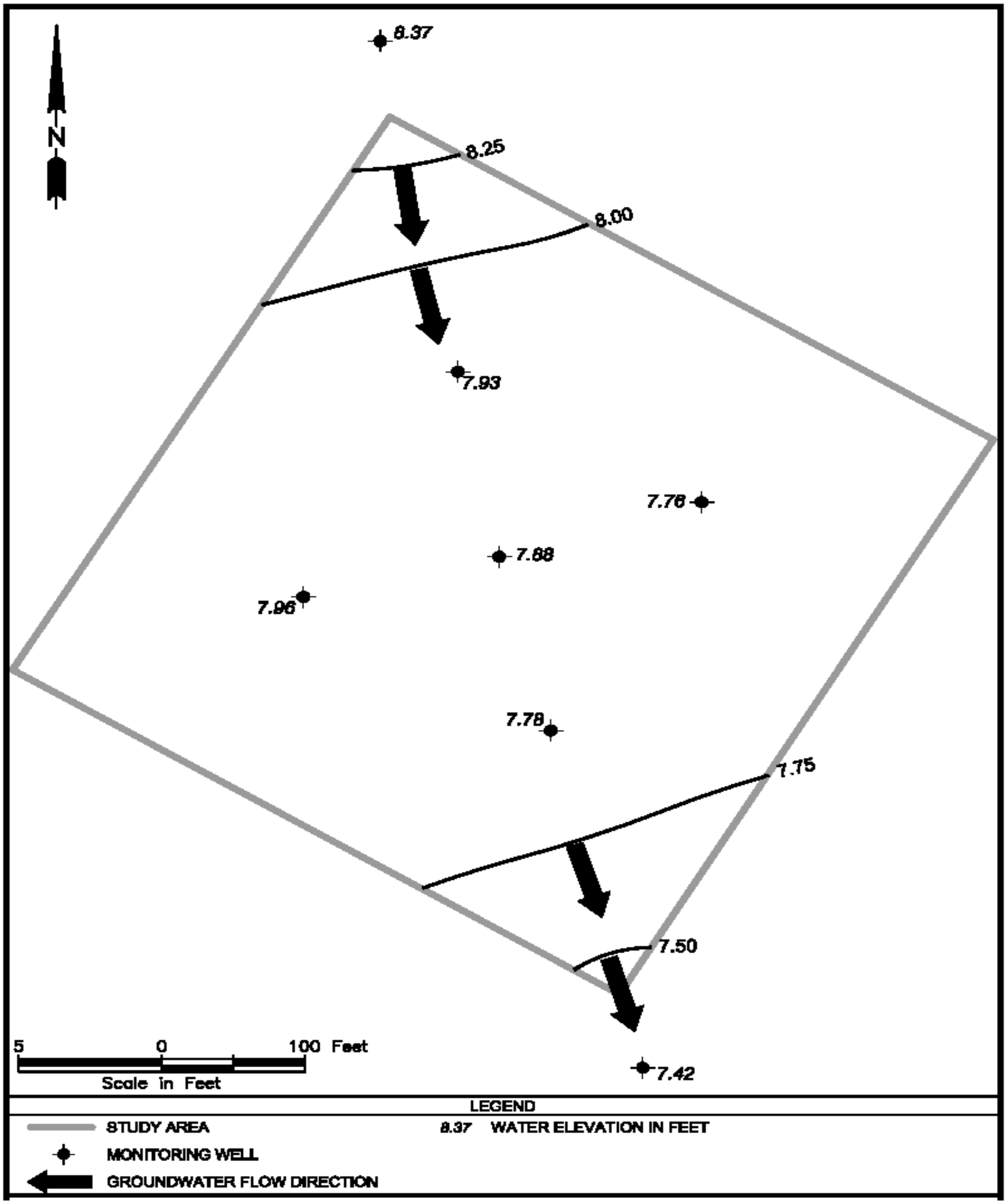


Figure J-24. Contour Plot Drawn by Modeling Software Using IDW.

APPENDIX K

Intervals and Limits

K-1. Introduction. Statistics can be divided into two categories: estimation theory (descriptive statistics) and hypothesis testing (inferential statistics). Estimation theory includes calculating confidence intervals as estimates for population parameters, while hypothesis testing focuses on the use of statistical tests to accept or reject hypotheses concerning these parameters.

K-2. Types of Statistical Intervals. Three types of statistical intervals are often constructed on data: confidence intervals, tolerance intervals, and prediction intervals. A confidence interval is designed to contain the specified population parameter, such as the mean, with a specified level of confidence. A confidence interval for the mean, for example, gives information about the average concentration level but offers little information about the highest or most extreme sample concentrations that are likely to be observed. In such cases, tolerance or prediction intervals are more appropriate. A confidence interval contains a parameter of interest, while a tolerance interval contains a proportion of the population, and a prediction interval contains one or more future observations. Statistical intervals are dependent upon distributional assumptions. Parametric and nonparametric methods for deriving intervals are also available. However, some nonparametric intervals, such as the tolerance interval, require a large number of observations to provide a reasonable coverage and confidence level. More information about statistical intervals can be found in Hahn and Meeker (1991). It should also be noted that the statistical software package ProUCL (Version 4.1) can be used to readily calculate most of parametric and non-parametric confidence, prediction and tolerance limits described in this Appendix. ProUCL can also be used to process data sets that contain non-detects. EPA freely distributes the software with a User's Guide, which can be downloaded from the website: http://www.epa.gov/osp/hstl/tsc/software.htm

K-2.1. Confidence Interval. It is often desirable to express or quantify the degree of uncertainty for some estimate of an unknown population parameter. The most common type of interval estimate is a confidence interval. A confidence interval is essentially an estimate for an unknown population parameter expressed as a range of values with some specified level of confidence. The level of confidence describes the probability that the "interval will capture the true parameter in repeated samples" (Moore, 1999).

K-2.1.1. The values at each end of the interval are called confidence limits. The lower value is the lower confidence limit (LCL) and the upper value is the upper confidence limit (UCL). The calculation of a confidence limit depends on the sampling distribution. Confidence limits are readily calculated for normally distributed data. A two-sided confidence interval for some population parameter, Θ, will be a closed interval of the form $a \leq \Theta \leq b$, where a is the lower limit and b is the upper limit. An upper one-sided confidence interval will be of the form $\Theta \leq b$ and a lower one-sided confidence interval will be of the form $\Theta \geq a$.

K-2.1.2. For environmental work, it is often desirable to estimate the mean concentration of a contaminant in some environmental population (for example, the mean concentration of arsenic in a shallow groundwater aquifer). The population mean (μ) is often estimated by calculating the sample mean ($\bar{x}$) for a set of *n* measurements. The uncertainty associated with the sample mean (as an estimate of the population mean) would be addressed by constructing a confidence interval for the population mean. A note on terminology: one calculates a confidence interval for a population parameter, such as the population mean, and *not* for the corresponding sample statistic, such as the sample mean (though a statistic such as the sample mean may be used to calculate the confidence interval for the population parameter).

K-2.1.3. The upper bound of the confidence interval of the population mean, the UCL, is most frequently encountered. For example, risk assessments require the 95% UCL for use as the reasonable maximum exposure concentration. The UCL of the (population) mean is used for the exposure point concentration (EPC) in risk assessments because of the uncertainty associated with estimating the population or "true" mean concentration at a site (EPA OB92-963373). Recent EPA guidance directs risk assessors in the possible methods used to calculate an upper confidence limit on the population mean (EPA OSWER 9285.6-10).

K-2.1.4. The phrase "95% confidence interval" means that "if one repeatedly calculates such intervals from many sets of independent random samples," 95% of the intervals, "in the long run, correctly contain the parameter of interest" (Hahn and Meeker, 1991). In other words, if a very large number of 95% confidence limits are calculated for the population mean, approximately 95% of the intervals (95 intervals out of 100) will contain the population mean. "More commonly, but less precisely, a two-sided confidence interval is described by a statement such as 'we are 95% confident that the interval contains the parameter of interest.' In fact, either the observed interval contains the parameter or it does not. Thus the 95% refers to the procedure for constructing a statistical interval, and not to the observed interval itself" (Hahn and Meeker, 1991). Because not all data sets fit a normal distribution, formulas for calculating a lognormal and nonparametric confidence limit are also available.

K-2.1.5. The EPA recently published OSWER 9285.6-10. According to this latest guidance, calculating a UCL should take into consideration outliers, censored data, and distribution testing (as described in Appendices I, H, and F). Once the distribution is determined, the calculation of an UCL should proceed according to the procedures for distributional methods. If, however, the site data do not follow a known distribution, then determining a good estimate of the UCL is left to the discretion of the risk assessor. Table K-1 presents the methods recommended in EPA guidance (OSWER 9285.6-10). Research in the area of UCL calculation is ongoing and recommendations may change in the future.

K-2.2. Tolerance Interval. A tolerance interval is designed to contain a specified proportion of the population (or percentile), such as 95% of all possible sample measurements (i.e., the 95th percentile). Tolerance intervals are essentially confidence intervals around a specified percentile. It is rare that a quantile for the population is known; instead, it is estimated using a sample data set, and a confidence interval for the population quantile is calculated using the sample quantile (e.g., just as a confidence interval for the population mean is calculated using the sample mean). Tolerance intervals are usually designed to cover all but a small percentage of the population measurements, so observations should rarely exceed a tolerance interval if the observations come from a similar distribution.

K-2.2.1. A tolerance interval is characterized by two quantities (probabilities): the coverage (the proportion of the population that the interval is supposed to contain), and the confidence level (the degree of confidence with which the interval reaches the specified coverage). As the interval is constructed from sample information, it is also a random interval. Because of sample fluctuations, a tolerance interval can contain the specified proportion of the population only with a certain confidence level. For example, "the $(1 - \alpha)100\%$ tolerance interval with $p100\%$ coverage" refers to a tolerance interval constructed to contain at least $100p\%$ of the distribution with $(1 - \alpha)100\%$ level of confidence.

K-2.2.2. Upper tolerance limits (UTLs) (UCLs for percentiles) are often calculated for environmental work. For example, it may be desirable to compare contaminant concentrations in a study area to the UTL of the compound in a background area. If the concentrations of many site samples exceed the background UTL, site-related contamination probably exists. It is most common for environmental scientists to calculate the "95 UTL" (95% upper tolerance limit with 95% coverage).

K-2.2.3. The method for calculating a tolerance interval depends on the nature of the underlying population distribution. Tolerance intervals can be constructed assuming that the data or the transformed data are normally distributed. It is also possible to construct nonparametric tolerance intervals using only the assumption that the data come from some continuous population. However, nonparametric tolerance intervals often require a large number of observations to provide a reasonable coverage and are impractical to construct for small sets of data. The data set with which tolerance intervals are calculated should be inspected for outliers and tested for normality before selecting the tolerance interval approach.

K-2.3. Prediction Interval. A prediction interval is a statistical interval calculated to include one or more future observations from the same population with a specified confidence. A prediction interval calculated from some set of sample data is such that all of certain number of future measurements (k) from the same population will fall within the interval with some specified level of confidence. In other words, each k future observation is compared to the prediction interval. The interval is constructed to contain all k future observations with the stated confidence. If any future observation exceeds the prediction interval, this is statistically significant evidence of a change in conditions. The number of future observations to be collected, *k*, must be specified (i.e., known before calculating the prediction interval). It is desirable to calculate prediction intervals periodically, using the most recent data. (The EPA recommends at least yearly for groundwater analyses.) Concentrations of site contaminants are sometimes compared to background concentrations using prediction intervals. An upper prediction limit is calculated for the next *k* future observations using the background data set and the *k* site measurements are then compared to the upper prediction limit. If any of the *k* site measurements exceed the prediction limit, this suggests that the site concentrations are elevated with respect to background.

Table K-1.
UCL Method Flow Chart

Are data normal?	Yes →	Use Student's *t*
No ↓		
Are data lognormal?	Yes →	Use Land, Chebyshev (MVUE), or Student's *t* (with small variance and skewness)
No ↓		
Is another distribution appropriate?	Yes →	Use distribution-specific method (if available)
No ↓		
Is sample size large?	Yes →	Use Central Limit Theorem-Adjusted (with small variance and mild skewness) or Chebyshev
No ↓		
→		Use Chebyshev, Bootstrap Resampling, or Jackknife

K-2.3.1. Prediction intervals are used to achieve some desired tolerance for Type I error (i.e., false rejection of H_0) when the same statistical test is performed multiple times (e.g., *k* times). (Neither prediction nor tolerance intervals address Type II error.) For example, assume that the Type I error rate is α for falsely rejecting the null hypothesis, H_0, for some statistical test or comparison. Assume that *k* independent statistical tests or comparisons are performed, where α denotes the probability of a false rejection (Type I error rate) for each individual test or comparison. The Type I error for the set of *k* independent comparisons, α*, is the following:

$$\alpha^* = 1 - (1-\alpha)^k .$$

K-2.3.2. Consider a single statistical test comparing populations 1 and 2, where H_0 is rejected at a level of significance $\alpha = 0.05$. Now, suppose that three, rather than two, populations are to be compared to each other using the same α for each comparison; that is, populations 1 and 2, 2 and 3, and 1 and 3, are compared, where $\alpha = 0.05$ for each of the $k = 3$ comparisons. Assume that the three populations are identical and all the measurements are independent of one another. The probability of rejecting H_0 for at least one of the three populations (i.e., the false rejection rate for the set of three comparisons) is

$$\alpha^* = 1 - (1-\alpha)^3 = 1 - (0.95)^3 = 0.14.$$

K-2.3.3. Even though the false rejection rate for a single comparison is 0.05, the false rejection rate for the set of three comparisons is higher, 0.14. A larger false positive rate will be obtained when more than three different populations are being compared. Therefore, if a total false rejection rate of $\alpha = 0.05$ is desired, the false rejection rate for each comparison must be less than 0.05. In fact, it can be shown that if a total false rejection rate (also called the experiment wise error rate) of α is desired, then the false rejection rate, α^*, for each comparison should be approximately α/k:

$$\alpha = 1 - (1-\alpha^*)^k \approx 1 - (1-\alpha/k)^k \ .$$

This is called the Bonferroni approximation. For example, if $\alpha = 0.05$ and $k = 3$, then the Type I error for each individual comparison (α^*) must be approximately 0.05/3 = 0.0167. Note that

$$1 - (1-\alpha/k)^k = 1 - (0.05/3)^3 = 0.049 \approx 0.05.$$

K-2.3.4. Thus, a prediction interval for the next k measurements for the $(1 - \alpha)100\%$ level of confidence that uses the Bonferroni approximation will entail the use of individual comparison with Type I error of α/k. For example, for normally distributed data, the prediction interval for $(1 - \alpha)100\%$ confidence for the next k observations is obtained from the quantile of the Student's t-distribution $t_{1-\alpha/k}$ (e.g., rather than $t_{1-\alpha}$).

K-2.3.5. It should be noted that, in general, prediction and tolerance intervals are not the same thing. The difference between a tolerance and prediction limit is one of interpretation and probability. Given n measurements and a desired confidence level, a tolerance interval will have a certain coverage percentage. A tolerance interval is designed so that, with some level of confidence, a proportion p of future measurements will fall within the interval. Thus, a small proportion $1 - p$ of the measures may fall outside the tolerance interval. A prediction limit, on the other hand, is designed so that, with some level of confidence, all future measurements fall within the interval. In this sense, the prediction limit may be thought of as a 100% coverage tolerance limit for the next k future observations. Thus, upper prediction intervals are constructed when all future measurements must fall below some threshold value and tolerance intervals are typically constructed when only a large proportion of future measurements are required to exceed a threshold value.

K-3. Statistical Intervals Based on Normal Distribution.

K-3.1. Confidence Interval for the Mean. For a normal distribution, the one-sided $(1 - \alpha)100\%$ UCL for the population mean is computed using the sample mean and standard deviation, and the $(1 - \alpha)$ quantile of Student's t-distribution with $n - 1$ degrees of freedom:

$$\mathrm{UCL}_{1-\alpha} = \bar{x} + t_{1-\alpha,n-1}\left(s/\sqrt{n}\right) .$$

Quantiles of the Student's t-distribution for various degrees of freedom are provided in Appendix B, Table B-23. Student's t can also be obtained in Microsoft Excel with the formula $TINV(2\alpha, n-1)$, for a one-sided (upper) $(1-\alpha)100\%$ UCL for $n-1$ degrees of freedom. When data are normally distributed, or if there are more than 30 samples available, a normal two-sided or one-sided confidence interval for the population mean (μ) with $(1-\alpha)100\%$ level of confidence can be computed as directed in the Paragraph K-3.2. An example is provided in Paragraph K-3.3.

K-3.2. Directions for the Confidence Interval for the Mean (Normal Distribution) When the Population Standard Deviation is Unknown. Let $x_1, x_2, \ldots, x_n$ represent the n data points from a normal distribution. These could be either n individual samples or n composite samples consisting of k aliquots each.

K-3.2.1. Verify that data come from a normal distribution using tests presented in Appendices F and J such as the Shapiro-Wilk test (Paragraph F-3) and a normal probability plot (Paragraph J-5.5).

K-3.2.2. Calculate the sample mean, $\bar{x}$, and the standard deviation, s (Appendix D).

K-3.2.3. Use Table B-23 of Appendix B to find the critical value such that $(1-\alpha)100\%$ of the t-distribution with $v = n-1$ degrees of freedom (df) is below this value. For a one-sided confidence interval (when just a LCL or an UCL is to be calculated), the critical value is the percentile $t_{1-\alpha,v}$. For a two-sided confidence interval (when both a LCL and UCL are to be calculated), the critical value is $t_{1-\alpha/2,v}$.

K-3.2.4. For example, if a two-sided 95% confidence interval is estimated, where $\alpha = 0.05$ and $n = 16$, then $v = n-1 = 16-1 = 15$ and $t_{1-(0.05/2),15} = t_{0.975,15} = 2.131$. If a one-sided 95% confidence interval is estimated, where $\alpha = 0.05$ and $n = 16$, then $v = n-1 = 16-1 = 15$ and $t_{1-0.05,15} = t_{0.95,15} = 1.753$.

K-3.2.5. For one-side confidence intervals for the population mean (μ), the equations for estimating the upper confidence limit (UCL) and lower confidence limit (LCL) are as follows:

$$\text{UCL} = \bar{x} + t_{1-\alpha,\nu}\left(s/\sqrt{n}\right)$$

$$\text{LCL} = \bar{x} - t_{1-\alpha,\nu}\left(s/\sqrt{n}\right).$$

K-3.2.6. The corresponding one-sided confidence intervals for are as follows:

$$\left(-\infty,\ \bar{x} + t_{1-\alpha,\nu}\left(s/\sqrt{n}\right)\right)$$

$$\left(\bar{x} - t_{1-\alpha,\nu}\left(s/\sqrt{n}\right), +\infty\right).$$

K-3.2.7. The two-sided confidence interval for the population mean is as follows:

$$\bar{x} \pm t_{1-\alpha/2,\nu}\left(s/\sqrt{n}\right).$$

K-3.3. <u>Example of a Confidence Interval for the Mean (Normal Distribution)</u>. Suppose a one-sided 95% lower confidence interval is desired for the mean concentration of (total) chromium in subsurface (below 5 feet from ground surface) soil at Site A.

K-3.3.1. These are the same data used in Paragraph L-6.1.3 as an example of a one-sample *t*-test. In that example there was evidence that the average was greater than 2 and not less than 2. A similar conclusion can also be reached when confidence intervals are constructed and compared to the regulatory threshold of 2, as illustrated in this example.

K-3.3.2. The first step is to verify that the data follow a normal distribution. The Shapiro-Wilk test is performed with these data; this test shows evidence that the data follow a normal distribution because the test's *p* value was 0.8489 and is greater than 0.05.

K-3.3.3. The mean and standard deviation of the data were calculated:

$$\bar{x} = 4.619$$

$$s = 0.8980\,.$$

Note that:

$\alpha = 0.05$ (for the 95% level of confidence)

$$n = 36$$

and

$$v = n - 1 = 36 - 1 = 35.$$

Table K-2.
Example Data

Site A Sample Location	Top Depth of Sample (ft)	Bottom Depth of Sample	Chromium (Total) Concentration (mg/kg) (ft)	Site A Sample Location	Top Depth of Sample (ft)	Bottom Depth of Sample (ft)	Chromium (Total) Concentration (mg/kg)
EPC-SB01	9	10	2.95	EPC-SB07	9	10	5.1
EPC-SB01	14	15	5.17	EPC-SB07	14	15	4.94
EPC-SB01	19	20	4.8	EPC-SB07	19	20	4.76
EPC-SB02	9	10	4.53	EPC-SB08	9	10	4.62
EPC-SB02	14	15	4.01	EPC-SB08	14	15	4.72
EPC-SB02	19	20	5.91	EPC-SB08	19	20	4.73
EPC-SB03	9	10	3.96	EPC-SB09	9	10	3.21
EPC-SB03	14	15	4.81	EPC-SB09	14	15	4.14
EPC-SB03	19	20	5.27	EPC-SB09	19	20	4.85
EPC-SB04	9	10	5.99	EPC-SB10	9	10	4.25
EPC-SB04	14	15	4.6	EPC-SB10	14	15	5.09
EPC-SB04	19	20	5.51	EPC-SB10	19	20	3.68
EPC-SB05	9	10	4.72	EPC-SB11	9	10	5.12
EPC-SB05	14	15	3.56	EPC-SB11	14	15	6.6
EPC-SB05	19	20	4.22	EPC-SB11	19	20	6.19
EPC-SB06	9	10	3.91	EPC-SB12	9	10	3.15
EPC-SB06	14	15	5.81	EPC-SB12	14	15	4.11
EPC-SB06	19	20	4.48	EPC-SB12	19	20	2.8

K-3.3.4. Using Table B-23 of Appendix B and linear interpolation, we find the critical value to be 1.691.

$$t_{1-\alpha,v} = t_{0.95,35} = (1.697 + 1.684)/2 = 1.691.$$

The confidence interval is

$$\left(\left\{4.619 - 1.691\left(0.8980/\sqrt{36}\right)\right\},\ \infty\right) = \left(4.37, \infty\right) .$$

K-3.3.5. The confidence interval does not contain 2 (the lower confidence limit exceeds 2); therefore, this is evidence that the average is greater than 2, the regulatory threshold.

K-3.4. Tolerance Interval (Normal Distribution). A one-sided tolerance limit is an upper or a lower confidence limit of a percentile (or proportion). A one-sided upper tolerance limit (UTL) that is greater than at least p100% of the population with probability $(1 - \alpha)$100% is the $(1 - \alpha)$100% upper confidence limit for the $p100^{\text{th}}$ percentile of the population (Hahn, 1970). Similarly, a one-sided lower tolerance limit (LTL) that is less than at least p100% of the population with probability $(1 - \alpha)$100% is the $(1 - \alpha)$100% lower confidence limit for the $p100^{\text{th}}$ percentile of the population. However, two-sided tolerance intervals are not equivalent to two-sided confidence intervals of percentiles. "Tolerance limits differ from confidence intervals in that tolerance limits provide an interval within which at least a proportion q of the population lies, within probability $1 - \alpha$ or more that the stated interval does indeed 'contain' the proportion q of the population" (Conover, 1999). "Two-sided tolerance intervals are rarely used in environmental studies, perhaps because there are few applications that attempt to determine the location of a central proportion of data, with allowable exceedances at both high and low ends" (Helsel, 2005).

K-3.4.1. Directions for a Tolerance Interval (Normal Distribution). Let $x_1, x_2, \ldots, x_n$ represent the n data points from a normal distribution. These could be either n individual samples or n composite samples consisting of k aliquots each. A two-sided $(1 - \alpha)$100% tolerance interval to contain at least p100% of a normal distribution is denoted as $(x_L,\ x_U)$, where x_L is the lower tolerance limit and x_U is the upper tolerance limit.

K-3.4.1.1. Verify data come from a normal distribution using tests presented in Appendices F and J, such as the Shapiro-Wilk test (Paragraph F-3) and a normal probability plot (Paragraph J-5.5).

K-3.4.1.2. Calculate the sample mean, $\bar{x}$, and the standard deviation, s (Appendix D).

K-3.4.1.3. For a two-sided tolerance interval, (x_L, x_U):

$$x_L = \bar{x} - s\, g_{1-\alpha,p,n}$$

$$x_U = \bar{x} + s\, g_{1-\alpha,p,n} \ .$$

K-3.4.1.4. Use Table B-14 of Appendix B to find the critical value g.

K-3.4.1.5. An approximation for g that may be useful (e.g., to find values of g that are not in Table B-13) is:

$$g_{1-\alpha,p,n} \approx Z_{(1+p)/2}\left(\frac{n-1}{\chi^2_{\alpha,n-1}}\right)^{1/2}\left(1+\frac{1}{2n}\right) .$$

Percentiles of the chi-square distribution, $\chi^2_{p,\nu}$, are listed in Table B-2. Percentiles of the standard normal distribution, Z_p, are listed in Table B-15. Hahn states that the approximation "appears to be good for most practical purposes even for *n* as small as 5" (Hahn, 1970).

K-3.4.1.6. For a one-sided lower tolerance limit, x_L:

$$x_L = \bar{x} - s\, g'_{1-\alpha,p,n} .$$

K-3.4.1.7. For a one-sided upper tolerance limit, x_U:

$$x_U = \bar{x} + s\, g'_{1-\alpha,p,n} .$$

K-3.4.1.8. Use Table B-13 of Appendix B to find the critical value *g´* (for values of *p* > 0.5).

K-3.4.1.9. An approximation for *g´* that may be useful is:

$$g'_{1-\alpha,p,n} \approx \frac{Z_p + \left(Z_p^2 - ab\right)^{1/2}}{a}$$

$$a = 1 - \frac{Z^2_{1-\alpha}}{2(n-1)}$$

$$b = Z_p^2 - \frac{Z^2_{1-\alpha}}{n} .$$

However, Hahn states that this approximation "is poor for very small *n*, especially for large *p* and large 1- α, and is not advised for *n* < 8" $-\alpha$ (Hahn, 1970).

K-3.4.2. <u>Example of a Two-sided Tolerance Interval (Normal Distribution)</u>. Suppose a two-sided 95% tolerance interval to contain at least 90% of the population is desired for chromium concentrations (total) in subsurface (below 5 feet from ground surface) soil at Site A, using the same data as Paragraph K-3.2.

K-3.4.2.1. The first step is to verify that the data follow a normal distribution. The Shapiro-Wilk test is performed with these data. This test shows evidence that the data follow a normal distribution because the test's *p* value was 0.8489 and is greater than 0.05.

K-3.4.2.2. The mean and standard deviation of the data were calculated:

$$\bar{x} = 4.619$$

$$s = 0.8980\,.$$

Note that:

$$p = 0.90$$

$$n = 36$$

$$\alpha = 0.05\,.$$

K-3.4.2.3. From Table B-14, $g_{0.95,0.90,35} = 2.090$ and $g_{0.95,0.90,40} = 2.052$. Therefore,

$$g_{0.95,0.90,36} = 2.090 - \frac{36-35}{40-35}(2.090 - 2.052) = 2.082\,.$$

K-3.4.2.4. The equation in Paragraph K-3.4.1.5 can also be used to calculate *g:*

$$g_{1-\alpha,p,n} \approx Z_{(1+0.90)/2}\left(\frac{36-1}{\chi^2_{0.05,35}}\right)^{1/2}\left(1+\frac{1}{2\times 36}\right) = 1.645\times\left(\frac{35}{22.46}\right)^{1/2}\times 1.014 = 2.082\,.$$

K-3.4.2.5. The two-sided tolerance interval is:

$$4.619 \pm 2.082 \times 0.8980 \text{ mg/kg}$$

(2.749, 6.489) mg/kg.

K-3.4.3. <u>Example of a One-Sided Upper Tolerance Limit, UTL (Normal Distribution)</u>. Suppose a UTL for the 95th percentile and 95% confidence level (also called a 95 UTL) is desired for chromium concentrations (total) in subsurface (below 5 feet from ground surface) soil at Site A, using the same data in Paragraph K-3.3.

K-3.4.3.1. As shown in the previous examples, the data seem to follow a normal distribution. For this example:

$p = 0.95$

$n = 36$

$\alpha = 0.05$

$1 - \alpha = 0.95$

$\bar{x} = 4.619$

$s = 0.8980$.

K-3.4.3.2. Using Table B-13 of Appendix B and linear interpolation, we find the critical value for the one-sided upper confidence limit to be

$$g'_{1-\alpha,\,p,\,n} = g'_{0.95,\,0.95,\,36} = 2.167 - \frac{36-35}{40-35}(2.167 - 2.125) = 2.159.$$

K-3.4.3.3. The approximation for *g´* in Paragraph K-3.4.1.9 may also be used to estimate *g´*:

$$a = 1 - \frac{Z^2_{1-\alpha}}{2(n-1)} = 1 - \frac{Z^2_{0.95}}{2(36-1)} = 1 - \frac{1.645^2}{70} = 0.9613$$
$$b = Z^2_p - \frac{Z^2_{1-\alpha}}{n} = Z^2_{0.95} - \frac{Z^2_{0.95}}{36} = 1.645^2 - \frac{1.645^2}{36} = 2.631$$

$$g'_{1-\alpha,p,n} \approx \frac{Z_p + \left(Z^2_p - ab\right)^{1/2}}{a} = \frac{1.645 + \left(1.645^2 - 0.9613 \times 2.631\right)^{1/2}}{0.9613} = 2.149 .$$

K-3.4.3.4. So, using the value for *g´* from Table B-13, the UTL is:

$$\text{UTL} = 4.619 + 0.8980 \times 2.159 = 6.558 \text{ mg/kg}.$$

K-3.4.4. Confidence Interval for the Variance or Standard Deviation (Normal Distribution). To estimate the precision of variance estimates, a confidence interval for the variance or standard deviation can be constructed. This information may be necessary for a sensitivity analysis of the statistical test or analysis method. The method described below can be used to find a two-sided $(1-\alpha)100\%$ confidence 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. Directions are presented in Paragraph K-3.4.4.1, followed by an example in Paragraph K-3.4.4.2.

K-3.4.4.1. Directions for a Confidence Interval for the Variance and Standard Deviation (Normal Distribution). Let $x_1, x_2, \ldots, x_n$ represent the *n* data points from a normal distribution.

K-3.4.4.1.1. Verify data come from a normal distribution using tests presented in Appendices F and J such as the Shapiro-Wilk test (Paragraph F-3) and a normal probability plot (Paragraph J-5.5).

K-3.4.4.1.2. Calculate the sample variance, s^2 (Appendix D).

K-3.4.4.1.3. For a $(1-\alpha)100\%$ two-sided confidence interval, use Table B-2 of Appendix B to find the critical values $\chi^2_{\alpha/2,v}$ and $\chi^2_{1-\alpha/2,v}$ with degrees of freedom *v* = (*n* – 1).

K-3.4.4.1.4. A $(1-\alpha)100\%$ confidence interval for the true underlying variance is $(s_L^2,\ s_U^2)$:

$$s_L^2 = \frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,v}}$$

$$s_U^2 = \frac{(n-1)\,s^2}{\chi^2_{\alpha/2,v}}.$$

K-3.4.4.1.5. A $(1-\alpha)100\%$ confidence interval for the true underlying standard deviation is (s_L, s_U):

$$s_L = \sqrt{\frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,v}}}$$

$$s_U = \sqrt{\frac{(n-1)\,s^2}{\chi^2_{\alpha/2,v}}}.$$

K-3.4.4.2. Example of Constructing a Confidence Interval for the Sample Variance and Standard Deviation (Normal Distribution). Consider the following data, background subsurface chromium concentrations of 3.84, 4.26, 4.53, 4.60, 5.28, 5.29, 5.74, and 5.86 mg/kg.

K-3.4.4.2.1. A confidence interval for the sample variance will be calculated based on a 95% level of confidence.

K-3.4.4.2.2. Testing the data for normality using the Shapiro-Wilk test indicated that the data were normal. So, a confidence interval for the sample variance based on a normal distribution can be calculated.

K-3.4.4.2.3. The sample variance, $s^2 = 0.526$. The required critical values are obtained from Table B-2:

$$\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,7} = 1.69$$

$$\chi^2_{(1-\alpha/2),n-1} = \chi^2_{0.975,7} = 16.01.$$

K-3.4.4.2.4. A 95% confidence interval for the true underlying variance is (0.228, 2.18):

$$s_L^2 = \frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,n-1}} = \frac{(8-1)(0.526)}{16.01} = 0.228$$

$$s_U^2 = \frac{(n-1)\,s^2}{\chi^2_{\alpha/2,n-1}} = \frac{(8-1)(0.526)}{1.69} = 2.18.$$

K-3.4.4.2.5. A 95% confidence interval for the true underlying standard deviation is (0.479, 1.48):

$$s_L = \sqrt{\frac{(n-1)\,s^2}{\chi^2_{(1-\alpha/2),n-1}}} = \sqrt{0.228} = 0.479$$

$$s_U = \sqrt{\frac{(n-1)\,s^2}{\chi^2_{\alpha/2,n-1}}} = \sqrt{2.18} = 1.48.$$

K-3.4.5. Prediction Interval (Normal Distribution). The prediction interval presented here is constructed assuming that the data follow a normal distribution with unknown mean and standard deviation. Most evaluations with environmental data only need a one-sided prediction interval, so this discussion will focus on the one-sided, upper prediction limit. To obtain a two-sided prediction interval, first replace α by $\alpha/2$. Then use the equation for the upper limit as the lower limit after replacing the addition of the standard deviation term with subtraction. The prediction interval must specify the overall level of confidence. That means a prediction interval's confidence level must account for the level of confidence of every future comparison. This is accomplished by setting the confidence level for each of the *k* future comparisons to $(1-\alpha/k)100\%$. Directions for calculating an upper prediction limit are presented in Paragraph K-3.4.5.1, followed by an example in Paragraph K-3.4.5.2.

K-3.4.5.1. Directions for Calculating an Upper Prediction Limit for k Future Comparisons of the Mean Calculated from m Observations (Normal Distribution). Verify the assumptions of normality.

K-3.4.5.1.1. The population mean and standard deviation are unknown. Specify *k* and *m* for the interval, where the mean of *m* observations is taken *k* times in the future (i.e., *k* samples are analyzed and the result reported for each sample is the mean of *m* replicate measurements).

K-3.4.5.1.2. Specify the level of confidence for the upper prediction limit as $(1-\alpha)100\%$.

K-3.4.5.1.3. Calculate the upper prediction limit

$$\bar{x} + s\,t_{1-\alpha/k,\,n-1}\sqrt{\frac{1}{m}+\frac{1}{n}}$$

where $\bar{x}$ is the mean of the original data, *s* is the standard deviation, and *n* is the total number of observations (measurements of the original data set).

K-3.4.5.1.4. Table B-23 of Appendix B provides values for $t_{1-\alpha/k,\,n-1}$.

K-3.4.5.1.5. If the future observations are found to be in the prediction interval, this is evidence that there has been no change in the sample values. If a future observation falls outside of the prediction interval, this is statistical evidence that the new observation does not come from the same distribution.

K-3.4.5.1.6. When replicate sample analyses are not done (i.e., a signal measurement or analysis is performed for each sample), set *m* = 1. For a single future observation (i.e., one sample analyzed once), set $m = 1$ and $k = 1$.

K-3.4.5.2. Example of Calculating a Normal Upper Prediction Limit for k Future Comparisons of the Mean from m Observations (Normal Distribution). A prediction interval is calculated for a set of "background well" measurements to determine if a set of "compliance well" measurements are "elevated" relative to background levels. The background well data set was tested for normality using the Shapiro-Wilk test. Because the data set was not normally distributed, the data set was normalized by taking the natural logarithm of each result.

K-3.4.5.2.1. For the compliance well data set, $m = 4$ replicate measurements are made for *k* = 1 sample. Let $\alpha = 0.01$ for the prediction interval. For the background data set, *n* = 8.

$$\bar{x}+s\,t_{1-\alpha/k,\,n-1}\sqrt{\frac{1}{m}+\frac{1}{n}}=-4.317+0.2832\,(2.998)\sqrt{\frac{1}{4}+\frac{1}{8}}=-3.79$$

$$t_{1-\alpha/k,\,n-1}=t_{1-\frac{0.01}{1},8-1}=t_{0.99,\,7}=2.998$$

using Table B-23 of Appendix B.

Table K-3.
Example Compliance Well Data

Background Well	Sample Date	Result	Log Result	Compliance Well	Sample Date	Result	Log Result
69-2-07	2001	0.0137	–4.290	69-2-08	2002	0.563	–0.574
69-2-07	2001	0.019	–3.963	69-2-08	2002	0.512	–0.669
69-2-07	2001	0.0163	–4.117	69-2-08	2002	0.475	–0.744
69-2-07	2001	0.0195	–3.937	69-2-08	2002	0.546	–0.605
69-2-07	2001	0.0112	–4.492				
69-2-07	2001	0.0112	–4.492				
69-2-07	2001	0.0102	–4.585				
69-2-07	2001	0.00946	–4.661				
Mean	-	0.01382	–4.317			0.524	–0.6484
Std. Dev.	-	0.00398	0.2832			0.0389	0.0753

K-3.4.5.2.2. Because the data set was transformed by taking the natural logarithm prior to calculating the upper prediction limit, to express the calculated limit in terms of the original units, it is necessary to perform the inverse transformation (i.e., to take the exponent of the calculated limit): exp(–3.79) = 0.0226. Therefore, the prediction interval is (0, 0.0226). Now we can compare the mean of the compliance well observations (0.524) with the upper limit of the prediction interval (0.022) calculated from the background well data. As 0.524 > 0.0226, there is significant evidence that the compliance well observations do not come from the same distribution as the background well.

K-4. Statistical Intervals Based on Lognormal Distribution.

K-4.1. Confidence Interval for the Mean.

K-4.1.1. When data are truly lognormal, it is not recommended that confidence intervals be calculated using the natural-log transformed data and the normal confidence intervals. One reason is that the units as well as the confidence intervals would be in log scale. The confidence intervals cannot be transformed back to the original scale and original units without a special adjustment.

K-4.1.2. For a lognormal distribution (the second alternative provided in the EPA UCL method flow chart), the EPA recommends calculating the UCL of the mean using one of several options based on the sample size, *n*, and the standard deviation of the log-transformed data, s_y. Table K-4, which summarizes these recommendations, has been adapted from the ProUCL Version 3.0 User Guide (EPA 600/R-97/006). However, it should be noted that the ProUCL Version 4.1 User Guide (EPA 600/R-07/041) is the most recent guidance. ProUCL Version 4.1 differs from Version 3.0 with respect to how UCLs are calculated when data include non-detects. ProUCL Version 4.1 uses non-parametric methods that appropriately treat non-detects as inequalities, while Version 3.0 allows only surrogate numerical values to be assigned to the non-detects (e.g., increasing the uncertainty of the UCL calculations). (Refer to Appendix H for additional information.) In addition to the computational methods listed below, the most current version of the software uses the gamma distribution to calculate UCLs. The software calculates UCLs using a number of different computational methods and automatically selects the "best" method (e.g., using criteria similar to that presented in Table K-4). However, it should be noted that these computational methods can result in relatively large UCLs (e.g., near the maximum detected values when the distributions are extremely skewed). This problem can be potentially avoided or at least minimized by collecting composite rather than grab samples (when possible and consistent with data quality objectives), as this tends to normalize data (i.e., composite samples produced from a sufficiently large number of grabs tend to be normally distributed).

K-4.2. Land Method.

K-4.2.1. Introduction. The Land method was touted in older EPA guidance, but it is no longer recommended in all cases because it is very sensitive to deviations from lognormality. Recall that distribution tests are primarily tests that the fit assumption cannot be rejected, rather than that the fit is perfect. Consequently, it is possible to pass a test for lognormality even when there are deviations from that distribution. This outcome is more likely for small data sets (< 30), which are quite common in environmental applications. The UCL for the Land method is as follows:

$$\text{UCL}_{1-\alpha} = \exp\left(\bar{y} + \frac{s_y^2}{2} + \frac{H_{1-\alpha} s_y^2}{\sqrt{n-1}} \right) .$$

K-4.2.1.1. The value of the *H* statistic is available in some statistical texts, including Gilbert (1987) and in Table B-8 of Appendix B.

K-4.2.1.2. Directions for constructing a confidence interval for the population mean of a lognormal distribution using the Land method are given in Paragraph K-4.2.2, followed by an example in Paragraph K-4.2.3 (EPA 600/R-97/006).

Table K-4.
Recommended Methods for Computation of a 95% UCL for the Unknown Mean of a Lognormal Population

Standard Deviation of Log-Transformed Data, s_y	Sample Size, *n*	Recommended Method (Paragraph Reference)
$s_y < 0.5$	For all *n*	Student's *t* (K-3.4.4) or Land (K-4.1)
$0.5 \le s_y < 1.0$	For all *n*	Land (K-4.1)
$1.0 \le s_y < 1.5$	$n < 25$	95% Chebyshev (MVUE) UCL (K-4.1)
	$n \ge 25$	Land (K-4.1)
$1.5 \le s_y < 2.0$	$n < 20$	99% Chebyshev (MVUE) UCL (K-4.1)
	$20 \le n < 50$	95% Chebyshev (MVUE) UCL (K-4.1)
	$n \ge 50$	Land (K-4.1)
$2.0 \le s_y < 2.5$	$n < 20$	99% Chebyshev (MVUE) UCL (K-4.1)
	$20 \le n < 50$	97.5% Chebyshev (MVUE) UCL (K-4.1)
	$50 \le n < 70$	95% Chebyshev (MVUE) UCL (K-4.1)
	$n \ge 70$	Land (K-4.1)
$2.5 \le s_y < 3.0$	$n < 30$	Larger of (99% Chebyshev (MVUE) UCL (K-4.1) or 99% Chebyshev (Mean, Sd) (K-5)
	$30 \le n < 70$	97.5% Chebyshev (MVUE) UCL (K-4.1)
	$70 \le n < 100$	95% Chebyshev (MVUE) UCL (K-4.1)
	$n \ge 100$	Land (K-4.1)
$3.0 \le s_y < 3.5$	$n < 15$	Hall's Bootstrap* (K-4.1)
	$15 \le n < 50$	Larger of (99% Chebyshev (MVUE) UCL (K-4.1) or 99% Chebyshev (Mean, Sd) (K-5)
	$50 \le n < 100$	97.5% Chebyshev (MVUE) UCL (K-4.1)
	$100 \le n < 150$	95% Chebyshev (MVUE) UCL (K-4.1)
	$n \ge 150$	Land (K-4.1)
$s_y \ge 3.5$	For all *n*	Use non-parametric methods* (K-5)

*In case Hall's Bootstrap method yields an erratic unrealistically large UCL value, then the UCL of the mean may be computed based upon the Chebyshev inequality.

K-4.2.2. Directions for a Confidence Interval for the Mean (Lognormal Distribution, Land Method). Let $x_1, x_2, \ldots, x_n$ represent the *n* data points from a lognormal distribution.

K-4.2.2.1. Verify that data come from a lognormal distribution using tests presented in Appendices F and J such as the Shapiro-Wilk test (Paragraph F-3) and a normal probability plot (Paragraph J-5.5).

K-4.2.2.2. Using the log-transformed data, $y_i = Ln(x_i)$, calculate the sample mean, $\bar{y}$, and the standard deviation, s_y.

K-4.2.2.3. Use Table B-8 of Appendix B to find the critical value (also called the *H* statistic) for the given level of confidence, sample size, and standard deviation. If a two-sided confidence interval for the mean is desired (LCL, UCL), the critical values are $H_{\alpha/2,n,s_y}$ and $H_{1-\alpha/2,n,s_y}$ for the LCL and UCL, respectively. If a one-sided confidence interval for the mean is desired, the critical value for the LCL is H_{α,n,s_y}, or the critical value for an UCL is $H_{1-\alpha,n,s_y}$. To estimate *H* values not in the table, a four-point Lagrangian interpolation (cubic interpolation) should be implemented.

K-4.2.2.4. For a two-sided confidence interval for the mean, the equations are as follows:

$$\mathrm{LCL} = \exp\left(\bar{y} + \frac{s_y^2}{2} + \frac{s_y H_{\alpha/2}}{\sqrt{n-1}}\right), \quad \mathrm{UCL} = \exp\left(\bar{y} + \frac{s_y^2}{2} + \frac{s_y H_{1-a/2}}{\sqrt{n-1}}\right).$$

K-4.2.2.5. For a one-sided confidence interval for the mean, *LCL* or *UCL*, the equation is as follows:

$$\mathrm{LCL} = \exp\left(\bar{y} + \frac{s_y^2}{2} + \frac{s_y H_{\alpha}}{\sqrt{n-1}}\right) \text{ or } \mathrm{UCL} = \exp\left(\bar{y} + \frac{s_y^2}{2} + \frac{s_y H_{1-a}}{\sqrt{n-1}}\right).$$

K-4.2.3. Example of a Confidence Interval for the Mean (Lognormal Distribution), Land Method. Suppose a one-sided 95% UCL is desired for concentrations of chromium (total) in background subsurface soil (5 feet below ground surface).

Sample ID	**Result (mg/kg)**	***Ln*(Result) (*Ln* mg/kg)**
EPC-BG01-013	0.0196	–3.932
EPC-BG01-020	0.00605	–5.108
EPC-BG02-010	0.00485	–5.329
EPC-BG02-020	0.0101	–4.595
EPC-BG03-010	0.00756	–4.885
EPC-BG03-020	0.00596	–5.123
EPC-BG04-010	0.0143	–4.248
EPC-BG04-020	0.00499	–5.300
EPC-BG05-010	0.00997	–4.608
EPC-BG05-020	0.00464	–5.373
EPC-BG06-010	0.00813	–4.812
EPC-BG06-023	0.00313	–5.767
EPC-BG07-010	0.00834	–4.787
EPC-BG07-020	0.00579	–5.151
EPC-BG08-010	0.00638	–5.055
EPC-BG08-020	0.00517	–5.265

K-4.2.3. Example of a Confidence Interval for the Mean (Lognormal Distribution), Land Method. Suppose a one-sided 95% UCL is desired for concentrations of chromium (total) in background subsurface soil (5 feet below ground surface).

K-4.2.3.1. The first step is to verify that the data follow a lognormal distribution. The Shapiro-Wilk test was performed with the log-transformed data. This test shows evidence that the data follow a normal distribution because the test's *p* value was 0.6570 and is greater than 0.05.

K-4.2.3.2. Using the log-transformed data,

$$\bar{y} = -4.959$$

and

$$s_y = 0.4574.$$

K-4.2.3.3. The critical value is $H_{0.95,16,0.4574} = 2.007$. A four-point Lagrangian interpolation (cubic interpolation) was implemented to obtain this critical value. K-4.2.4 shows how the critical value $H_{0.95,16,0.4574}$ was derived.

K-4.2.3.4. For a one-sided upper confidence interval for the mean, UCL, the equation is:

$$\text{UCL} = \exp\left(\bar{y} + \frac{s_y^2}{2} + \frac{s_y H_{1-a}}{\sqrt{n-1}}\right) = \exp\left(-4.959 + \frac{0.4574^2}{2} + \frac{0.4574(2.007)}{\sqrt{16-1}}\right) = 0.0099.$$

K-4.2.4. Lagrangian Interpolation (Cubic Interpolation) for the H Statistic. The details of the Lagrangian (cubic) interpolation are provided to assist in the use of Table B-8 of Appendix B.

K-4.2.4.1. Suppose the *H* statistic $\left(H_{1-\alpha/2,n,s_y}\right)$ is desired for

$$1 - \alpha/2 = 0.95$$

$$n = 16$$

$$s_y = 0.4574$$

(from Paragraph K-4.2.3).

K-4.2.4.2. A Lagrangian interpolation requires bounding the desired value by two tabulated values lower and two tabulated values higher than the desired value. Using the example above, we need a column of *H* statistics when $n = 16$ because there is no such column in Table B-8. The tabulated columns *n* = 12, 15 (two values below 16) and *n* = 21, 31 (two values above 16) are used to generate a column for $n = 16$. Once the column of *H* statistics is generated for $n = 16$, Lagrangian interpolation can be used to get the *H* statistic for $s_y = 0.4574$.

K-4.2.4.3. So the columns associated with s_y = 0.30, 0.40 (two values below 0.4574) and s_y = 0.50, 0.60 (two values above 0.4574) are used to generate a column for s_y = 0.4574.

K-4.2.4.4. From Table B-8, the following *H* statistics, $H_{0.95,n,s_y}$, are needed for these interpolations:

			n		
s_y	12	15	16	21	31
0.30	1.927	1.882	$H_{0.95,16,0.30}$	1.833	1.793
0.40	2.026	1.968	$H_{0.95,16,0.40}$	1.905	1.856
0.4574	—	—	$H_{0.95,16,0.4574}$	—	—
0.50	2.141	2.068	$H_{0.95,16,0.50}$	1.989	1.928
0.60	2.271	2.181	$H_{0.95,16,0.60}$	2.085	2.010

K-4.2.4.5. The first part of the interpolation process is to generate a column of *H* statistics for $n = 16$. For each s_y, the following equation is used:

$$H_{0.95,16,s_y} = \frac{(16-15)(16-21)(16-31)}{(12-15)(12-21)(12-31)} H_{0.95,12,s_y} + \frac{(16-12)(16-21)(16-31)}{(15-12)(15-21)(15-31)} H_{0.95,15,s_y} + \frac{(16-12)(16-15)(16-31)}{(21-12)(21-15)(21-31)} H_{0.95,21,s_y} + \frac{(16-12)(16-15)(16-21)}{(31-12)(31-15)(31-21)} H_{0.95,31,s_y}.$$

So,

$$\begin{aligned} H_{0.95,16,0.30} &= \frac{(16-15)(16-21)(16-31)}{(12-15)(12-21)(12-31)}(1.927) + \frac{(16-12)(16-21)(16-31)}{(15-12)(15-21)(15-31)}(1.882) + \\ &\quad \frac{(16-12)(16-15)(16-31)}{(21-12)(21-15)(21-31)}(1.833) + \frac{(16-12)(16-15)(16-21)}{(31-12)(31-15)(31-21)}(1.793) \\ &= -0.2817 + 1.960 + 0.2037 - 0.0118 = 1.8702. \end{aligned}$$

The same process was used to determine $H_{0.95,16,0.40}$, $H_{0.95,16,0.50}$, and $H_{0.95,16,0.60}$.

n

s_y	12	15	16	21	31
0.30	1.927	1.882	1.870	1.833	1.793
0.40	2.026	1.968	1.953	1.905	1.856
0.4574	—	—	$H_{0.95,16,0.4574}$	—	—
0.50	2.141	2.068	2.049	1.989	1.928
0.60	2.271	2.181	2.158	2.085	2.010

K-4.2.4.6. Next, the H statistic values for the various s_y at $n = 16$ are used to interpolate $H_{0.95,16,0.4574}$.

$$
\begin{aligned}
H_{0.95,16,0.4574} &= \frac{(0.4574-0.40)(0.4574-0.50)(0.4574-0.60)}{(0.30-0.40)(0.30-0.50)(0.30-0.60)}(1.870) \\
&+ \frac{(0.4574-0.30)(0.4574-0.50)(0.4574-0.60)}{(0.40-0.30)(0.40-0.50)(0.40-0.60)}(1.953) \\
&+ \frac{(0.4574-0.30)(0.4574-0.40)(0.4574-0.60)}{(0.50-0.30)(0.50-0.40)(0.50-0.60)}(2.049) \\
&+ \frac{(0.4574-0.30)(0.4574-0.40)(0.4574-0.50)}{(0.60-0.30)(0.60-0.40)(0.60-0.50)}(2.158) \\
&= -0.1087 + 0.9337 + 1.320 - 0.1384 \\
&= 2.007.
\end{aligned}
$$

Thus, the H statistic is 2.007.

K-4.3. Chebyshev (MVUE) Method.

K-4.3.1. Introduction. For the Chebyshev (MVUE) method, first estimate the mean and variance using the minimum unbiased variance approach discussed in Appendix D. Then calculate the $100(1-\alpha)\%$ UCL of the mean using:

$$
\mathrm{UCL}_{1-\alpha} = \hat{\mu}_1 + \sqrt{\left(\frac{1}{\alpha}-1\right)s^2\left(\hat{\mu}_1\right)}\ .
$$

The quantities $\hat{\mu}_1$ and $s^2\left(\hat{\mu}_1\right)$are the MVUE estimates of the mean and standard deviation given in equations D-2 and D-3 in Appendix D. An example of using this method follows in Paragraph K-4.3.2.

K-4.3.2. Example of a Confidence Interval for the Mean (Lognormal Distribution), Chebyshev MVUE Method. Suppose chromium concentrations (mg/kg) measured at a site are as follows:

0.378	1.411	1.089	0.918
0.073	0.518	2.240	0.111
1.246	2.251	1.967	1.894
1.414	13.844	1.222	0.962
0.094	0.247	0.371	0.056

K-4.3.2.1. ProUCL was used to determine the 95% UCL for the population mean. The data follow a lognormal distribution (Shapiro-Wilk *p* = 0.905 on the log-transformed data). The sample size is 20, and the standard deviation of the log-transformed values is 1.39. Table J-2 recommends using the 95% Chebyshev MVUE UCL as the 95% UCL for the population mean under these conditions.

K-4.3.2.2. The MVUE estimate of the mean, $\hat{\mu}_1$, is 1.66, and the standard deviation of the estimate of the mean, $s^2\left(\hat{\mu}_1\right)$, is 0.607. Therefore,

$$\text{UCL}_{0.95} = \hat{\mu}_1 + \sqrt{\left(\frac{1}{\alpha} - 1\right)s^2\left(\hat{\mu}_1\right)} = 1.66 + \sqrt{\left(\frac{1}{0.05} - 1\right)\left(0.607\right)^2} = 4.30\,.$$

K-4.4. Hall’s Bootstrap Method.

Although Hall’s Bootstrap is a nonparametric method related to the Bootstrap technique presented in Paragraph K-5, a limited presentation will be given here because EPA guidance (OSWER 9285.6-10) specifically recommends this technique for calculating the UCL of a lognormal population under certain situations described in Table K-4. The method adjusts for bias and skewness in the data (OSWER 9285.6-10). Directions for implementing Hall’s Bootstrap are given in Paragraph K-4.4.1 and results of Hall’s method from ProUCL Version 3.0 are presented in Paragraph K-4.4.2. The directions for performing the bootstrap method are presented for illustration only, as bootstrap methods require too many arithmetic calculations for manual calculations to be practical.

K-4.4.1. Directions for Implementing Hall’s Bootstrap Method for a: $100(1-\alpha)\%$ UCL. Let $x_1, x_2, \ldots, x_n$ represent *n* randomly sampled concentrations.

K-4.4.1.1. Compute the sample mean,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,.$$

K-4.4.1.2. Compute the sample standard deviation,

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\ .$$

K-4.4.1.3. Compute the sample skewness,

$$k = \frac{1}{ns^3}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3\ .$$

K-4.4.1.4. Do the following a large number of times:

K-4.4.1.4.1. Generate a simple random sample of *n* values from $x_1, x_2, \ldots, x_n$ with replacement.

K-4.4.1.4.2. Compute the sample mean, $\bar{x}_i$, standard deviation, s_i, and skewness, k_i, of the sample found in K-4.4.1.4.1.

K-4.4.1.4.3. Compute the Studentized mean,

$$W_i = \frac{\left(\bar{x}_i - \bar{x}\right)}{s_i}.$$

K-4.4.1.4.4. Compute Hall's statistic,

$$Q_i = W_i + \frac{k_i\, W_i^2}{3} + \frac{k_i^2\, W_i^3}{27} + \frac{k_i}{6n}.$$

K-4.4.1.4.5. Sort all the values, Q_i, in ascending sequence and calculate the α^{th} lower quantile, Q_α.

K-4.4.1.4.6. Calculate

$$W = \frac{3}{k}\left\{\left[1 + k\left(Q_\alpha - \frac{k}{6n}\right)\right]^{\frac{1}{3}} - 1\right\}\ .$$

The one-sided $(1-\alpha)100\%$ upper confidence limit is $UCL_{1-\alpha} = \bar{x} - Ws$.

K-4.4.2. <u>Example of a Confidence Interval for the Mean (Lognormal Distribution) Hall's Bootstrap</u>. Suppose chromium concentrations (mg/kg) measured at a site are as follows:

0.331	0.104
68.977	0.022
0.908	2.044
140.605	0.093
157.359	0.213

ProUCL was used to determine the 95% UCL for the population mean. The data follow a lognormal distribution (Shapiro-Wilk p = 0.842 on the log-transformed data). The sample size is 10, and the standard deviation of the log-transformed values is 3.27. Table K-4 recommends using Hall's Bootstrap to estimate the 95% UCL for the population mean under these conditions. The Bootstrap algorithm was run with the result $\mathrm{UCL}_{0.95}$ = 71.4 mg/kg. Because this result is based on random sampling, it may change with repeated runs. As a comparison, the Land method 95% UCL for this data is over 3,240,000 mg/kg (an unrealistically large value).

K-4.4.3. Confidence Interval for a Percentile–Tolerance Interval (Lognormal Distribution). A lognormal confidence interval for the $p100^{\text{th}}$ percentile of a lognormal distribution, X_p, with $(1-\alpha)100\%$ confidence, can be derived by using the log-transformed data with the equations for the normal confidence interval. When $Y = Ln(X)$ is normal (i.e., X is lognormal), given a set of sample values $y_1, y_2 \ldots y_n$ with sample mean $\bar{y}$ and standard deviation s, the exponent of $\bar{y}$ is an estimate of the 50th percentile (median) of X ($X_{0.5}$):

$$x_{0.5} = \exp(\bar{y}) \ .$$

K-4.4.3.1. The two-sided $100(1-\alpha)\%$ confidence interval for the median of X is:

$$\left(\exp\left(\bar{y} - \frac{t_{\alpha/2,n-1}\, s}{\sqrt{n}}\right), \exp\left(\bar{y} + \frac{t_{1-\alpha/2,n-1}\, s}{\sqrt{n}}\right)\right) .$$

K-4.4.3.2. In general, if X is lognormal and $Y = Ln(X)$, then an estimate x_p of the $p100^{\text{th}}$ percentile of X (X_p) is obtained by first calculating an estimate of Y_p (the $p100^{\text{th}}$ percentile of Y),

$$y_p = \bar{y} + t_{p,n-1}\, s$$

and then performing the inverse transformation (exponentiation) on this quantity. The (maximum likelihood) estimate of the percentile X_p in terms of the original variable (X) is:

$$x_p = \exp(y_p) = \exp\left(\bar{y} + t_{p,n-1}\, s\right).$$

K-4.4.3.3. A one-sided upper confidence limit for the percentile X_p is calculated as follows:

$$\exp(\bar{y} + g'_{1-\alpha,p,n} s) \text{ for } p > 0.5 \text{ .}$$

K-4.4.3.4. The term in parentheses is simply a confidence limit for a normal percentile or tolerance limit as described in Paragraph K-3.4.

K-4.4.3.5. A two-sided tolerance interval is calculated as follows:

$$\left(\exp(\bar{y} - g_{1-\alpha,p,n} s), \exp(\bar{y} + g_{1-\alpha,p,n} s\right) \text{ .}$$

K-4.4.4. Prediction Interval (Lognormal Distribution). A lognormal prediction interval can be calculated using the log-transformed data with the process for developing normal prediction intervals. When *X* is lognormal and *Y = Ln(X)* with sample mean $\bar{y}$ and standard deviation *s*, then the prediction interval for the next *k* observations in the original scale is:

$$\left(\exp\left(\bar{y} - \frac{t_{\alpha/2k,n-1}\, s}{\sqrt{1+1/n}} \right), \exp\left(\bar{y} + \frac{t_{1-\alpha/2k,n-1}\, s}{\sqrt{1+1/n}} \right) \right) \text{ .}$$

K-5. Distribution-Free Statistical Intervals.

K-5.1. Introduction. The one-sided Chebyshev inequality for a mean can be used when no distribution can be assumed to fit the data. Regardless of the underlying probability distribution of some variable *X*, the following inequality will be satisfied for the $(1 - \alpha)$100% UCL of the population mean μ:

$$(1-\alpha)100\% \text{ UCL} \le \bar{x} + \sqrt{\frac{1}{\alpha} - 1} \left(\frac{\sigma}{\sqrt{n}} \right) \text{ .}$$

K-5.1.1. The right-hand side of the inequality serves as a conservative estimate of the UCL. However, as the population standard deviation σ is typically unknown, the UCL is usually estimated as follows:

$$(1-\alpha)100\% \text{ UCL} \approx \bar{x} + \sqrt{\frac{1}{\alpha} - 1} \left(\frac{s}{\sqrt{n}} \right) \text{ .}$$

K-5.1.2. Unfortunately, because the sample standard deviation population (*s*) is being used to estimate the population standard deviation (σ), the population mean may not actually be less than this limit at the prescribed level of confidence when the variance or skewness is large, especially for small sample sizes. See OSWER 9285.6-10 for more details.

K-5.1.3. This one-sided Chebyshev UCL, based on the mean and standard deviation, is recommended for use with the lognormal distribution under certain conditions described in Table K-4. In that situation use the untransformed data to calculate $\bar{x}$ and *s*.

K-5.2. Confidence Interval for the Mean. If data do not follow either a normal or lognormal distribution, EPA guidance (OSWER 9285.6-10) recommends using either the central limit theorem or Bootstrap resampling. Several methods are available for estimating confidence limits of the mean when no distributional assumptions are made. The Bootstrap and Jackknife procedures are nonparametric statistical techniques that can be used to construct approximate confidence intervals for parameters such as the population mean. These procedures are nonparametric or distribution-free because they do not require assumptions about the data's distribution (such as normal or lognormal). It should be noted that statistical methods that account for the data's distribution, when used appropriately, are more efficient than the nonparametric methods. Directions for the Bootstrap and Jackknife methods for estimating a nonparametric confidence interval for θ, the parameter of interest, are given in Paragraphs K-5.2.1 and K-5.2.2, respectively. Examples are presented in Paragraphs K-5.2.3 and K-5.2.4. It should be noted that the both the Bootstrap and Jackknife methods are usually performed using statistical software owing to the large number of manual calculations that would be required. The Paragraphs below illustrate how the calculations are done.

K-5.2.1. Directions for a Bootstrap Estimate of the Confidence Interval for θ. Let $x_1, x_2, \ldots, x_n$ be a random sample of size *n*.

K-5.2.1.1. The parameter of interest is θ and a reasonable estimate of θ is $\hat{\theta}$. For example, θ is the mean and $\hat{\theta}$ is the minimum variance unbiased estimator (MVUE) of the mean (Appendix D).

K-5.2.1.2. Take *n* samples with replacement from the original set of random samples of size *n*, and define this new set of data as $x_{11}, x_{12}, \ldots, x_{1n}$. Note that the same result can be selected more than once. For this new data set, estimate $\hat{\theta}$ and denote it as $\hat{\theta}_1$.

K-5.2.1.3. Perform the previous step *N* times, each time calculating an estimate of $\hat{\theta}$. Denote all *N* estimates of $\hat{\theta}$ as $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N$. *N* should be considerably larger, such as 1000 or more. It is much easier to perform this simulation using a computer.

K-5.2.1.4. Estimate the Bootstrap estimate of $\theta, \bar{\theta}_B$, from the *N* estimates of $\hat{\theta}_I$, such that

$$\bar{\theta}_B = \frac{1}{N}\sum_{I=1}^{N}\hat{\theta}_I$$

for $i = 1, 2, \ldots, N$.

K-5.2.1.5. Derive the confidence interval for θ, with $100(1-\alpha)\%$ level of confidence, as (θ_L, θ_U) where $\theta_L = (\alpha/2)100^{th}$ percentile from the set of *N* estimates and $\theta_U = (1-\alpha/2)100^{th}$ percentile from the set of *N* estimates (see Appendix G). A one-sided UCL is simply the $(1-\alpha)100^{th}$ percentile from the set of *N* estimates.

K-5.2.2. Example of the Bootstrap Method for Estimating a Nonparametric Confidence Interval for θ. A confidence interval for the population mean (μ) will be calculated for chromium concentrations in subsurface soil at Site A with 95% level of confidence. All chromium concentrations were detected so no proxy concentrations are needed to evaluate the data.

K-5.2.2.1. The data are as follows: 2.95, 5.17, 4.80, 4.53, 4.01, 5.91, 3.96, 4.81, 5.27, 5.99, 4.60, 5.51, 4.72, 3.56, 4.22, 3.91, 5.81, 4.48, 5.10, 4.94, 4.76, 4.62, 4.72, 4.73, 3.21, 4.14, 4.85, 4.25, 5.09, 3.68, 5.12, 6.60, 6.19, 3.15, 4.11, and 2.80 mg/kg.

K-5.2.2.2. An example of 10 samples with replacement taken from the original set of random samples of size *n* = 36 is as follows: 2.95, 5.17, 5.91, 3.96, 4.80, 4.81, 4.53, 5.27, 4.01, and 5.99 mg/kg. (Note that although replacement was adhered to, no sample's values were actually "picked" twice.)

K-5.2.2.3. For this new data set, estimated mean is $\hat{\theta}_1 = 4.74$.

K-5.2.2.4. Perform the previous step *N* times, and each time calculating an estimate of $\hat{\theta}$. Using a statistical software package,

$$\bar{\theta}_B = \frac{1}{N}\sum_{I=1}^{N}\hat{\theta}_I = 4.626$$

for $i = 1, 2, \ldots, N$.

K-5.2.2.5. The confidence interval for θ, with 95% level of confidence reached upon 12 repetitions, is 4.323 to 4.93.

K-5.2.3. Directions for a Jackknife Estimate of the Confidence Interval for θ. Estimate $\hat{\theta}$ with all *n* samples from the data set.

K-5.2.3.1. Estimate $\hat{\theta}_{(i)}$ by removing the *i*th sample (for $i = 1, 2, \ldots, n$) from the original data set and use the same equation as was used to estimate $\hat{\theta}$.

K-5.2.3.2. Estimate the arithmetic mean, $\tilde{\theta}$, from the *n* estimates of $\hat{\theta}_{(i)}$, such that

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$$

for $i = 1,2,\ldots,n$. Note that the i^{th} "pseudo-value" is defined as $J_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$.

K-5.2.3.3. Calculate the Jackknife estimator of θ (the average of the J_i values),

$$J\left(\hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} J_i = n\hat{\theta} - (n-1)\,\tilde{\theta}\ .$$

K-5.2.3.4. Estimate the standard error of the Jackknife estimate, $J\left(\hat{\theta}\right)$, by

$$\hat{\sigma}_{J\left(\hat{\theta}\right)} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(J_i - J\left(\hat{\theta}\right)\right)^2}\ .$$

K-5.2.3.5. Derive the confidence interval as

$$\left(J\left(\hat{\theta}\right) - t_{1-(\alpha/2),n-1}\,\hat{\sigma}_{J\left(\hat{\theta}\right)},\ J\left(\hat{\theta}\right) + t_{1-(\alpha/2),n-1}\,\hat{\sigma}_{J\left(\hat{\theta}\right)}\right)$$

with $100(1-\alpha)\%$ level of confidence; $t_{p,n-1}$ is the critical value from the Student's *t*-distribution for the *p100*th percentile and *n* – 1 degrees of freedom. If only a one-sided confidence interval is needed, then $t_{p,n-1} = t_{1-\alpha,n-1}$.

K-5.2.4. Example of the Jackknife Method for Estimating a Nonparametric Confidence Interval for θ. Using the same data set as for the Bootstrap example (Paragraph K-5.2.1), we will calculate a confidence interval for the mean (μ) using the Jackknife estimate with a 95% level of confidence.

K-5.2.4.1. Estimate $\hat{\theta} = 4.62$ with all 36 samples from the data set.

K-5.2.4.2. Estimate $\hat{\theta}_{(i)}$ for *i* = 1, 2 … *n* = 36. The results are listed in Table K-5.

K-5.2.4.3. Estimate the arithmetic mean,

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} = 4.75\,.$$

K-5.2.4.4. Calculate the Jackknife estimator of θ (the average of the J_i values),

$$J\left(\hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} J_i = n\hat{\theta} - (n-1)\tilde{\theta} = 4.62 .$$

K-5.2.4.5. Estimate the standard error of the Jackknife estimate, $J\left(\hat{\theta}\right)$, by

$$\hat{\sigma}_{J\left(\hat{\theta}\right)} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(J_i - J\left(\hat{\theta}\right)\right)^2} = 0.15 .$$

K-5.2.4.6. Derive the confidence interval,

$$J\left(\hat{\theta}\right) \pm t_{1-(\alpha/2),n-1}\,\hat{\sigma}_{J\left(\hat{\theta}\right)}$$

$$J\left(\hat{\theta}\right) - t_{1-(\alpha/2),n-1}\,\hat{\sigma}_{J\left(\hat{\theta}\right)} = 4.62 - 1.69 \times 0.15 = 4.37$$

$$J\left(\hat{\theta}\right) + t_{1-(\alpha/2),n-1}\,\hat{\sigma}_{J\left(\hat{\theta}\right)} = 4.62 + 1.69 \times 0.15 = 4.87 .$$

K-5.2.4.7. The critical value from the Student's *t*-distribution was found using Table B-23 in Appendix B and linear interpolation.

K-5.3. Tolerance and Prediction Intervals. An approximate two-sided nonparametric prediction interval to contain the next single observation from the population with $(1-\alpha)100\%$ confidence can be estimated from the sample as $\left(x_{(l)},\, x_{(u)}\right)$ where

$$l = \frac{\alpha}{2}(n+1)$$

$$u = \left(1 - \frac{\alpha}{2}\right)(n+1)$$

and $x_{(i)}$ is the i^{th} order statistic from the sample data (Helsel and Hirsch, 2003). If l or u is not an integer, linearly interpolate between the values of the two surrounding order statistics. One-sided prediction limits can be calculated by replacing $\alpha/2$ with α when calculating the order statistic to use. An example calculation follows in Paragraph K-5.3.1.

K-5.3.1. Example of a One-Sided Nonparametric Prediction Limit for the Next Single Observation. A 95% upper prediction limit for arsenic concentration at a single point in the future is desired. Arsenic concentrations at three background wells were measured once each month for 12 months to yield 36 observations. Of the 36 observations, 19 were non-detects, so a nonparametric prediction limit will be calculated. The 95% upper prediction limit is calculated as:

$$x_{(u)} = x_{([1-\alpha][n+1])} = x_{([1-0.05][36+1])} = x_{(35.15)} \; .$$

Because 35.15 is not an integer, interpolate between the 35th and 36th order statistics. Suppose $x_{(35)}$=12 ppb and $x_{(36)}$=13 ppb. Then the 95% upper prediction limit is estimated to be:

$$x_{(35)} + 0.15\left(x_{(36)} - x_{(35)}\right) = 12 + 0.15(13-12) = 12.15\,\text{ppb}.$$

If the result of the next observation were 8 ppb, we could conclude that arsenic concentration has not increased with 95% confidence.

K-5.3.2. <u>Discussion</u>. Exact confidence for using various order statistics from a sample to create nonparametric prediction intervals and limits can be calculated using the methods described in Hall et al. (1975). Their calculations expand to cover prediction intervals to contain *k* of *m* future observations instead of just a single future observation.

K-5.3.2.1. For small datasets, the method presented in Paragraph K-5.3.1 will require an order statistic that is smaller than the smallest observation in the dataset (for a minimum) or larger than the largest (for a maximum). In this situation, a nonparametric UTL or UPL is typically constructed using the minimum or maximum value of the set of observations. With high probability, the tolerance interval is designed to miss only a small percentage of the observations that arise from the same population as the data used to develop the tolerance limit. The coverage probability for the tolerance interval can be reported as either a minimum or an average value because, typically, we can only specify that the coverage probability of the interval exceed some level of confidence. We will use the average value. Given *n* measurements, using the maximum measurement as the UTL yields an average confidence of

$$\frac{n}{n+1}100\% \, .$$

K-5.3.2.3. A prediction limit involves the confidence probability associated with predicting that the next single observation will fall below the upper prediction limit, and is the same as the expected (mean) coverage of a similarly constructed UTL. Note that this is a special case for nonparametric prediction limits for the next single observation, not a general result. Furthermore, it can be shown that the probability of having *k* future samples all fall below the upper nonparametric prediction limit is $(1-\alpha) = n/(n+k)$ (i.e., the maximum value is the $[n/(n+k)]100\%$ upper prediction limit for *k* future measurements). Table B-11 in Appendix B lists these confidence levels for various choices of *n* and *k*. The false positive rate associated with a single prediction limit can be computed as one minus the confidence level. An example calculation follows in Paragraph K-5.3.3.

Table K-5.
Estimate of $\hat{\theta}_{(i)}$ for i = 1, 2 … *n* = 36

i	Mean $\hat{\theta}$	J_i	$J_i - J(\hat{\theta})$	$(J_i - J(\hat{\theta}))^2$
1	4.67	2.8	–1.82	3.31
2	4.67	2.95	–1.67	2.78
3	4.66	3.15	–1.47	2.16
4	4.66	3.21	–1.41	1.98
5	4.65	3.56	–1.06	1.12
6	4.65	3.68	–0.94	0.88
7	4.64	3.91	–0.71	0.50
8	4.64	3.96	–0.66	0.43
9	4.64	4.01	–0.61	0.37
10	4.63	4.11	–0.51	0.26
11	4.63	4.14	–0.48	0.23
12	4.63	4.22	–0.40	0.16
13	4.63	4.25	–0.37	0.14
14	4.62	4.48	–0.14	0.02
15	4.62	4.53	–0.09	0.01
16	4.62	4.6	–0.02	0.00
17	4.62	4.62	0.00	0.00
18	4.62	4.72	0.10	0.01
19	4.62	4.72	0.10	0.01
20	4.62	4.73	0.11	0.01
21	4.61	4.76	0.14	0.02
22	4.61	4.8	0.18	0.03
23	4.61	4.81	0.19	0.04
24	4.61	4.85	0.23	0.05
25	4.61	4.94	0.32	0.10
26	4.61	5.09	0.47	0.22
27	4.60	5.1	0.48	0.23
28	4.60	5.12	0.50	0.25
29	4.60	5.17	0.55	0.30
30	4.60	5.27	0.65	0.42
31	4.59	5.51	0.89	0.79
32	4.58	5.81	1.19	1.42
33	4.58	5.91	1.29	1.67
34	4.58	5.99	1.37	1.88
35	4.57	6.19	1.57	2.47
36	4.56	6.6	1.98	3.93

K-5.3.2.4. Balancing the ease with which nonparametric upper prediction limits are constructed is the fact that, given fixed numbers of original samples and future sample values to be predicted, the maximum confidence level associated with the prediction limit is also fixed. To increase the level of confidence, the only choices are to: i) decrease the number of future values to be predicted at any testing period, or ii) increase the number of original samples used in the test. Table B-11 of Appendix B can be used along these lines to plan an appropriate sampling strategy so that the false positive rate can be minimized and the confidence probability maximized to a desired level.

K-5.3.3. Example of a Nonparametric Prediction Limit for the Next k Observations. A prediction limit for arsenic concentration at $k = 2$ points in the future is desired. Arsenic concentration at three background wells was measured once each month for 6 months to yield 18 observations. As 9 of the 18 observations were non-detects, a nonparametric prediction limit will be calculated. The maximum detected result was 12 ppb, so this will be used as the upper prediction limit. Because $n = 18$ and $k = 2$, the probability of both future observations falling below the upper prediction limit of 12 is

$$100\frac{n}{n+k}\% = 100\frac{18}{18+2}\% = 90\%\,.$$

Thus 12 ppb is a 90% upper prediction limit for two future observations. The results of the two future observations were 8 and 14 ppb. As one of the new observations exceeds 12 ppb, we can conclude that arsenic concentration has increased with 90% confidence.

K-5.4. Nonparametric Confidence Intervals for Percentiles. A nonparametric confidence interval is based on an actual sample result and does not rely on any distributional assumptions. The nonparametric confidence interval is generally wider and requires more data than the corresponding normal distribution interval, and so the parametric distribution intervals should be used whenever it is appropriate. When $n \leq 20$, the nonparametric confidence interval is calculated using the binomial distribution.

K-5.4.1. Given a set of measurements, $x_1, x_2, ...x_n$, to calculate a nonparametric confidence interval for the quantile X_p, it is necessary to first order the values of x_i so that $x_{(1)} < x_{(2)} <...< x_{(n)}$. Therefore, the smallest value of the data set is $x_{(1)}$ and the largest is $x_{(n)}$. (Note the distinction between x_1 and $x_{(1)}$; the former is the first measured value of the data set and the latter is the smallest value of the data set.) A two-sided nonparametric confidence interval for a quantile X_p will be of the form:

$$x_{(a)} \leq X_p \leq x_{(b)}$$

where the probability that X_p lies in the above interval is $1-\alpha$:

$$P(x_{(a)} \leq X_p \leq x_{(b)}) = 1-\alpha\,.$$

K-5.4.2. The a^{th} largest value $x_{(a)}$ and b^{th} largest value $x_{(b)}$ of the data set (i.e., the numerical values of *a* and *b* that satisfy the above equation) are determined using the binomial distribution (as will be discussed below). Unfortunately, because the values are selected from a finite set of *n* ordered values { $x_{(i)}$ }, confidence limits are essentially being constructed for a discrete rather than a continuous variable. In general it will not be possible to select *a* and *b* so that the above probability is exactly equal to $1-\alpha$. Therefore, for the two-sided $1-\alpha$ confidence interval, *a* and *b* are selected so that:

$$P(x_{(a)} \le X_p \le x_{(b)}) \ge 1-\alpha\,.$$

K-5.4.3. Similarly, for an upper one-sided confidence interval for a percentile X_p it is desirable to select *b* so that:

$$P(X_p \le x_{(b)}) \ge 1-\alpha.$$

Find the lower bound $x_{(a)}$ by selecting the value of *a* so that:

$$Bin(a-1,n,p) \le \alpha/2 \text{ and } Bin(a,n,p) > \alpha/2$$

where $Bin(k,n,p)$ denotes the probability for the cumulative binomial distribution—the probability that an event with probability *p* of occurrence will happen less than or equal to *k* times out of *n* trials:

$$Bin(k,n,p) = P(K \le k) = \sum_{i=0}^{k}\left[\frac{n!}{i!(n-i)!}p^i(1-p)^{n-i}\right].$$

K-5.4.4. More information on the binomial distribution can be found in Appendix F. The values of *n* and *p* are known. Table B-1 of Appendix B lists values of the cumulative binomial distribution and lists various values of *k* for fixed values of *p* and *n*. Because *p* (the quantile) and *n* (the number of samples) are known, we can use Table B-1 to find the appropriate value of *k*. For example, one could start with $k=0$, then $k=1$, and so forth until $k=a-1$ is the smallest value that satisfies the inequalities $Bin(a-1,n,p) \le \alpha/2$ and $Bin(a,n,p) > \alpha/2$. The upper bound $x_{(b)}$ is obtained by determining the smallest value of *b* that satisfies the relationship

$$Bin(b-1,n,p) - Bin(a-1,n,p) \ge 1-\alpha\,.$$

K-5.4.5. For example, let us calculate the two-sided nonparametric confidence limit for the 75th percentile (*p* = 0.75) for the 90% level of confidence ($\alpha = 0.1$) for $n = 16$ so that:

$$P(x_{(a)} \le X_{0.75} \le x_{(b)}) \ge 0.9.$$

From Table B-1,

$$Bin(8,16,0.75) = 0.0271 < \alpha/2 = 0.05$$

and

$$Bin(9,16,0.75) = 0.0796 > \alpha/2 = 0.05.$$

Therefore, $a = 9$. Because

$$Bin(14,16,0.75) - Bin(8,16,0.75) = 0.9365 - 0.0271 = 0.9094 > 1-\alpha = 0.9$$

the value for $b = k+1 = 14+1 = 15$. Therefore, the 90% confidence interval for the 75th percentile is $x_{(9)} \le X_{0.75} \le x_{(15)}$.

K-5.4.6. Similarly, find the one-sided $100(1-\alpha)\%$ upper confidence limit of X_p, so that the smallest value of *b* satisfies the equation

$$P(X_p \le x_{(b)}) = Bin(b-1, n, p) \ge 1-\alpha.$$

K-5.4.7. Once b is found from Table B-1, the b^{th} largest value, $x_{(b)}$, establishes the upper $(1-\alpha)100\%$ confidence limit of X_p. For example, if $n = 20$, $p = 0.5$, and $\alpha = 0.05$,

$$Bin(13,20,0.5) = 0.94 \text{ and } Bin(14,20,0.5) = 0.98.$$

K-5.4.8. Because $Bin(14,20,0.5) > 0.95$, $b = k+1 = 14+1 = 15$. The 15th largest value of the data set, $x_{(15)}$, is at least the 95% upper confidence limit of the 50th percentile: $P(X_{0.5} \le x_{(15)}) \ge 0.95$.

K-5.4.9. If $n > 20$, the two-sided $100(1-\alpha)\%$ confidence interval $x_{(a)} \le X_p \le x_{(b)}$ can be calculated using a normality approximation so that $P(x_{(a)} \le X_p \le x_{(b)}) \approx 1-\alpha$.

K-5.4.10. Calculate the following

$$a = np - Z_{1-\alpha/2}\sqrt{np(1-p)}$$

and

$$b = np + Z_{1-\alpha/2}\sqrt{np(1-p)}$$

where the percentile Z_p is the p^{th} quantile for the standard normal distribution obtained from Table B-15 of Appendix B. Round *a* an *b* to the nearest whole numbers and find the corresponding order values $x_{(a)}$ and $x_{(b)}$.

K-5.4.11. For the one-sided upper $100(1-\alpha)\%$ confidence interval $X_p \leq x_{(b)}$, where $P\left(X_p \leq x_{(b)}\right) \approx 1-\alpha$, calculate

$$b = np + Z_{1-\alpha}\sqrt{np(1-p)}\,.$$

Round to the nearest whole number and find $x_{(b)}$.

K-5.4.12. Maximum detected values can be used to make inferences about percentiles. In particular, assume that a set of detected values are ranked from lowest to highest so that $x_{(n)}$ denotes the maximum value. Also assume that the maximum detected value is less than some threshold concentration (i.e., a risk-based limit) *C*: $x_{(n)} < C$. It can be shown that, under these circumstances, if X_p is the *p*100th percentile of *X*, then

$$P(X_p \leq C) \geq 1 - p^n \text{ and } P(X_p > C) \leq p^n \ .$$

X_p is less than the threshold *C* with at least $1 - p^n = 1 - \alpha$ confidence.

K-5.4.13. To find the value of *n* needed to achieve the desired level of confidence $(1-\alpha)100\%$, *n* must be such that

$$p^n \leq \alpha \,.$$

Therefore, the *p*100th percentile, X_p, will be less the decision limit *C* with at least $(1-\alpha)100\%$ confidence if the maximum detected value is less than *C* (i.e., $x_{(n)} < C$) and

$$n \geq Ln(\alpha)/Ln(p)\,.$$

K-5.4.14. If, for example, $p = 0.90$ and $\alpha = 0.05$, then $n \geq 28.4$. If 29 samples are collected and the maximum value is less than *C*, then one can be at least 95% confident that the 90th percentile is less than *C*.

K-5.4.15. The maximum is a non-parametric one-sided upper tolerance limit. Given a set of *n* observed measurements, there is $(1 - \alpha)100\% = (1 - p^n)100\%$ confidence that at least *p*100% of future measurements will be less than the maximum. A two-sided tolerance interval to contain at least a proportion *p* of future measurements may be constructed using the minimum and maximum of a set of *n* observed measurements. There is

$$(1-\alpha)100\% = \left(1 - p^{n} - n(1-p)\,p^{n-1}\right)100\%$$

confidence that at least p100% of future measurements will fall between the minimum and maximum of set of n observed data points. For example, if n =50 and p = 0.95, then there is 72% confidence that at least 95% of future measurements will fall between the minimum and maximum.

K-6. Statistical Confidence Interval for Proportions. Data from a binomial distribution are composed of only two responses—"pass" or "fail." The population proportion, P, is based on either the passing proportion or the failing proportion. The following discussion will (arbitrarily) define the proportion, p, as the proportion of failures. An estimate of this proportion can be derived by $p = k/n$ where k is the number of failures out of n samples. For example, in environmental applications p could represent the proportion of results from samples below some decision limit, C. From this information we would like to estimate an interval, (P_L, P_U), which contains the true proportion, P, of the distribution that is less (or greater) than C. The binomial distribution is a discrete distribution and so statistical intervals are approximate and tend to be conservative (Hahn and Meeker, 1991). The most frequent statistical interval calculated for a proportion is the confidence interval, so only it is presented here.

K-6.1. Discussion. The equation for a conservative two-sided 100(1 – α) % confidence interval for a proportion is the following:

$$[p_L, p_U] = \left[\frac{1}{1+\left\{\dfrac{(n-k+1)\,F_{1-\alpha/2,\,2n-2k+2,\,2k}}{k}\right\}},\ \frac{1}{1+\left\{\dfrac{(n-k)}{(k+1)F_{1-\alpha/2,\,2k+2,\,2n-2k}}\right\}}\right]$$

where $F_{\gamma,m,n}$ is the $\gamma100^{\text{th}}$ percentile of the F distribution (Table B-7 of Appendix B) with m and n degrees of freedom. The lower limit, P_L, is defined to be 0 if k = 0, and the upper limit, P_U, is defined to be 1 if k = n (Hahn and Meeker, 1991).

K-6.1.1. Likewise, a one-sided $(1-\alpha)100\%$ LCL for a proportion would be:

$$p_L = \frac{1}{1+\left\{\dfrac{(n-k+1)\,F_{1-\alpha,\,2n-2k+2,\,2k}}{k}\right\}}$$

while a one-sided $(1-\alpha)100\%$ UCL for a proportion would be:

$$p_U = \frac{1}{1+\left\{\dfrac{(n-k)}{(k+1)F_{1-\alpha,\ 2k+2,\ 2n-2k}}\right\}} \ .$$

K-6.1.2. If a large number of samples are available, these confidence intervals can be approximated. However, two restrictions apply to the data set: first, $np \geq 5$ and second, $n(1 - p) \geq 5$. This approximated confidence interval is based on the normal distribution because when these two restrictions apply, data are approximately normally distributed. The equation for the approximated confidence interval is:

$$[p_L, p_U] = p \pm Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)100^{th}$ percentile from a standard normal, n is the sample size, and p is the sample proportion (Devore, 1987). The one-sided upper confidence limit would be found by replacing $1-\alpha/2$ with $1-\alpha$ as follows:

$$p_U = p + Z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}} \ .$$

K-6.2. Example of a Confidence Limit for a Proportion. Groundwater concentrations of gasoline at a site are compared to a regulatory threshold of 35 micrograms per liter (µg/L). Suppose out of 90 results, 11 of the samples have concentrations that exceed this regulatory threshold, so the proportion of samples with detected concentrations exceeding the threshold is $p = 11/90 = 0.1222$.

$$np = 90 \times 0.1222 = 11.00$$

$$n(1-p) = 90 \times (1-0.1222) = 79.00 .$$

As both np and $n(1 - p)$ are greater than or equal to 5, the large sample normal approximation can be used

$$p_L = p - Z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}} = 0.1222 - 1.282\sqrt{\frac{0.1222(1-0.1222)}{90}} = 0.078$$

where Table B-15 of Appendix B is used to find the critical value $Z_{0.90}$=1.282. Because p_L exceeds 0.05, we can accept that more than 5% of the concentrations of gasoline in groundwater at the site exceed the regulatory threshold as we conclude also in Appendix L, Paragraph L-8.2.

K-7. Statistical Intervals for the Poisson Distribution (Number of Occurrences). Data from a Poisson distribution are composed of only two mutually exclusive responses—"pass" or "fail" s—when the probability of one of the responses is small. Poisson distributions are common when counting the number of pass or fail occurrences over a time interval or the number of detections when a set of measured concentrations consists mostly of non-detects. The population rate of occurrence, μ, also called the mean rate of occurrence, is either based on the passing rate or the failing rate. For this document, the rate of occurrence of the rare event is called the rate of "failure." The Poisson distribution is a discrete distribution and so statistical intervals are approximate and tend to be conservative (Hahn and Meeker, 1991). An estimate of this rate of occurrence can be derived by:

$$\hat{\mu} = k / n$$

where *k* is the number of failures out of *n* samples.

K-7.1. Confidence Interval for the Mean Occurrence Rate. A two-sided $(1-\alpha)100\%$ confidence interval for the mean occurrence rate is the following:

$$[\hat{\mu}_L, \hat{\mu}_U] = \left[\frac{0.5\chi^2_{\alpha/2,\, 2k}}{n}, \frac{0.5\chi^2_{1-\alpha/2,\, 2k+2}}{n} \right]$$

where $\chi^2_{\gamma,v}$ is the $\gamma\,100^{\text{th}}$ percentile of the chi-square distribution (Table B-2 of Appendix B) with *v* degrees of freedom.

K-7.1.1. A one-sided lower or upper $(1-\alpha)100\%$ confidence limit can be obtained by replacing $\chi^2_{\alpha/2,2k}$ with $\chi^2_{\alpha,2k}$ for a lower confidence limit or replacing $\chi^2_{1-\alpha/2,2k+2}$ with $\chi^2_{1-\alpha,2k+2}$ for an upper confidence limit (Hahn and Meeker, 1991).

K-7.1.2. If a large number of samples is available (generally, if $n > 20$), this confidence interval can be approximated. This approximated confidence interval is based on the normal distribution because, as the sample size increases, the data's distribution tends towards normality. The equations for the approximated confidence interval are:

$$[\hat{\mu}_L, \hat{\mu}_U] = \hat{\mu} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}}{n}}$$

where $Z_{1-\alpha/2}$ is the $(1-\alpha/2)100^{th}$ percentile from a standard normal, *n* is the sample size, and $\hat{\mu}$ is the mean sample rate of failure ($\hat{\mu} = k/n$ when *k* is the number of failures in *n* samples) (Hahn and Meeker, 1991).

K-7.2. Upper Tolerance Limit. A Poisson tolerance interval, with p100% coverage and $(1-\alpha)100\%$ confidence, is calculated based on the directions given in Paragraph K-6.2.1, followed by an example in Paragraph K-6.2.2.

K-7.2.1. Directions for Calculating a Poisson Tolerance Interval with p100% Coverage and $(1-\alpha)100\%$ Confidence. Compute the sum of the Poisson counts of n samples:

$$T' = \sum_{i=1}^{n} x_i \ .$$

This is the sum of the detected values and one-half the sum of all the non-detected values.

K-7.2.1.1. Find the probable rate

$$\mu = \frac{1}{2n}\chi^2_{1-\alpha,2T'+2}$$

where $\chi^2_{1-\alpha,2T'+2}$ is the $(1-\alpha)100^{th}$ percentile of the chi-squared distribution with $\nu = 2T'+2$ degrees of freedom. Table B-2 of Appendix B contains a table of critical values for the chi-square distribution.

K-7.2.1.2. Compute the $p100^{\text{th}}$ percentile of the Poisson distribution with mean rate μ, by finding the least positive integer k such that

$$\chi^2_{1-p,2k+2} \geq 2\mu \, .$$

As above, the quantity $2k + 2$ represents the degrees of freedom of the chi-squared distribution. The quantity k itself is the upper tolerance limit (UTL) for the Poisson distribution. In other words, for the smallest value of k for which

$$\chi^2_{1-p,2k+2} \geq \frac{1}{n}\chi^2_{1-\alpha,2T'+2}$$

p100% of the measurements will be less than k with $(1-\alpha)100\%$ confidence. If any sample exceeds the UTL, k, then there is significant evidence that this sample is different from the samples used to develop the UTL.

K-7.2.2. Example of Calculating a Poisson Tolerance Interval with p100% Coverage and $(1-\alpha)100\%$ Confidence. A tolerance interval with 95% confidence (α = 0.05) and 95% coverage (p = 0.95) is desired for 1,1-dicholorethene in groundwater at Site B. The background well values in Table K-6 were obtained. These data have more than 90% non-detects and the number of samples n = 90.

K-7.2.2.1. Calculate the sum of the Poisson counts: Sum the detections and to this value add one half the sum of the non-detects (one half the detection limit is being used for each non-detect).

$$T' = 0.5 \times (7.1236) + (0.111 + 0.138 + 2.63 + 4.81) = 11.25$$

$$v = 2T' + 2 = 2\ (11.25) + 2 = 24.5 \approx 25$$

$$\mu = \frac{1}{2n}\chi^2_{1-\alpha,2T'+2} = \frac{1}{2 \times 90}\chi^2_{0.95,25} = 0.209$$

where $\chi^2_{1-\alpha,2T+2} = \chi^2_{0.95,25} = 37.65$ using Table B-2 of Appendix B.

K-7.2.2.2. So, we need to find the smallest value of *k* such that $\chi^2_{1-p,2k+2} \geq 2\mu$; that is, the value of *k* such that $\chi^2_{0.05,2k+2} \geq 0.418$. Table B-2 of Appendix B shows that the smallest value number of degrees of freedom, *v* = 2*k* + 2, that satisfies the above equation is *v* = 4. Since 4 = 2*k* + 2, *k* = 1.0.

k	df	$\chi^2_{0.005}$
0.5	3	0.3518
1	4	0.7107
1.5	5	1.145

K-7.2.2.3. If any site groundwater sample exceeds the UTL of 1.0 μg/L derived from the background wells, then there is significant evidence that contamination at the site is elevated with respect to background.

K-7.2.3. Upper Prediction Limit. To estimate a prediction limit using the Poisson model, the upper limit is estimated for an interval that will contain all of *k* future measurements of an analyte with $100(1-\alpha)\%$ confidence, given *n* previous measurements. The directions to calculate such a prediction limit are provided in Paragraph K-6.2.3.1 and followed by an example in Paragraph K-6.2.3.2.

K-7.2.3.1. Directions for Estimating a Prediction Limit Using the Poisson Model. Calculate T', the sum of the Poisson counts of *n* samples (e.g., for the background data set), as defined in Paragraph K-6.2.1.

K-7.2.3.1.1. Calculate T^*_k, the greatest total Poisson count for the next *k* samples (e.g., for the study area data set) at some level of confidence, 1 – α using the following equation:

$$T_k^* = \frac{T'}{n} + \frac{t^2}{2n} + \frac{t}{n}\sqrt{T'\left(1+n\right) + \frac{t^2}{4}}$$

where $t = t_{1-\alpha/k,n-1}$ is the upper $(1-\alpha/k)100\%$ percentile of the Student's *t*-distribution with *n* – 1 degrees of freedom, in Table B-23 of Appendix B.

K-7.2.3.1.2. If the sum of Poisson counts for the next *k* samples is greater than the upper prediction limit T_k^*, then there is significant evidence of a difference in the new samples, compared to previous samples.

K-7.2.3.2. Example of Estimating a Prediction Limit Using the Poisson Model. Suppose a prediction limit for the next two observations with 99% confidence is desired for 1,1-dicholorethene from Site B with the following background wells. NOTE: These data have more than 90% non-detects. (See data table in Paragraph K-6.2.2.)

K-7.2.3.2.1. Calculate the sum of the Poisson counts:

$$T' = 0.5\times(7.1236)+(0.111+0.138+2.63+4.81) = 11.25$$

$$T_k^* = \frac{T'}{n} + \frac{t^2}{2n} + \frac{t}{n}\sqrt{T'\left(1+n\right) + \frac{t^2}{4}}$$

$$= \frac{11.25}{90} + \frac{(2.639)^2}{2(90)} + \frac{2.639}{90}\sqrt{11.25\left(1+90\right) + \frac{(2.639)^2}{4}} = 1.10$$

where *n* = 90 and $t_{1-\alpha/k,n-1} = t_{(1-0.01)/2,(90-1)} = t_{0.995,89} = 2.639$ using Table B-23 of Appendix B and linear interpolation.

K-7.2.3.2.2. To test the upper prediction limit, if the sum of the Poisson counts for the next *k* samples (*k* = 2) is greater than T_k^* (1.10), then there is significant evidence the contamination in the site wells is elevated relative to the background wells.

Table K-6.
Background Wells

Well Location	Result (µg/L)	DL (µg/L)	Well Location	Result (µg/L)	DL (µg/L)
Site B-MW01		0.0819	SiteB-MW02		0.144
SiteB-MW01		0.102	SiteB-MW02		0.0715
SiteB-MW01		0.102	SiteB-MW02		0.0715
SiteB-MW01		0.0715	SiteB-MW02		0.145
SiteB-MW01		0.0436	SiteB-MW03		0.144
SiteB-MW01		0.0436	SiteB-MW03		0.0715
SiteB-MW01		0.122	SiteB-MW03		0.0715
SiteB-MW02		0.0819	SiteB-MW03		0.0715
SiteB-MW02		0.102	SiteB-MW04		0.144
SiteB-MW02		0.102	SiteB-MW04		0.0715
SiteB-MW02		0.0715	SiteB-MW04		0.0715
SiteB-MW02	0.111		SiteB-MW04		0.0715
SiteB-MW02		0.0436	SiteB-MW05		0.144
SiteB-MW02		0.122	SiteB-MW05		0.0715
SiteB-MW03		0.0819	SiteB-MW05		0.0715
SiteB-MW03		0.102	SiteB-MW05		0.0715
SiteB-MW03		0.102	SiteB-MW06		0.0715
SiteB-MW03		0.0715	SiteB-MW06		0.0715
SiteB-MW03		0.0436	SiteB-MW06		0.0715
SiteB-MW03		0.0436	SiteB-MW06		0.145
SiteB-MW03		0.122	SiteB-MW01		0.116
SiteB-MW04		0.0819	SiteB-MW01		0.116
SiteB-MW04		0.102	SiteB-MW01		0.0492
SiteB-MW04		0.102	SiteB-MW01		0.0492
SiteB-MW04		0.0715	SiteB-MW02		0.116
SiteB-MW04		0.0436	SiteB-MW02	0.138	
SiteB-MW04		0.0436	SiteB-MW02		0.0492
SiteB-MW04		0.122	SiteB-MW02		0.0492
SiteB-MW05		0.0819	SiteB-MW03		0.116
SiteB-MW05		0.102	SiteB-MW03		0.116
SiteB-MW05		0.102	SiteB-MW03		0.0492
SiteB-MW05		0.0715	SiteB-MW03		0.0492

Well Location	Result (µg/L)	DL (µg/L)	Well Location	Result (µg/L)	DL (µg/L)
SiteB-MW05		0.0436	SiteB-MW04		0.116
SiteB-MW05		0.0436	SiteB-MW04		0.116
SiteB-MW05		0.122	SiteB-MW04		0.0492
SiteB-MW06		0.0819	SiteB-MW04		0.0492
SiteB-MW06		0.102	SiteB-MW05		0.116
SiteB-MW06		0.102	SiteB-MW05		0.116
SiteB-MW06		0.0715	SiteB-MW05	2.63	
SiteB-MW06		0.0436	SiteB-MW05		0.0492
SiteB-MW06		0.0436	SiteB-MW06		0.116
SiteB-MW06		0.122	SiteB-MW06		0.116
SiteB-MW01		0.144	SiteB-MW06	4.81	
SiteB-MW01		0.0715	SiteB-MW06		0.0492
SiteB-MW01		0.0715			
SiteB-MW01		0.0715			

APPENDIX L

Hypothesis Testing—Simple Cases

L-1. Introduction. This Appendix provides an extensive discussion of the statement of hypotheses (null and alternative) and the consequences deriving from that choice. Also, a general introduction of the basic types of hypothesis testing commonly employed in environmental operations is provided. Further reading on the foundations of hypothesis testing can be found in EPA 600/R-96/055, QA/G-4. Additional reading on the one-sample hypothesis tests presented below can be found in EPA/240/B/026/003, QA/G-9S.

L-2. Translating Objectives into Statistical Hypotheses. A data user's question, or a decision rule from the DQO process, must be translated into a precise statistical statement to be tested using environmental data. Such a statement is called a hypothesis. It includes a null hypothesis (H_0) and an alternative hypothesis (H_A). The null hypothesis is a baseline condition presumed to be true in the absence of strong evidence to the contrary, and the alternative hypothesis is the opposite condition that bears the burden of proof. In other words, unless it is demonstrated that the alternative hypothesis is correct based upon weight of evidence, the baseline condition is retained.

L-2.1. A hypothesis test consists of the following elements.

L-2.1.1. It has a quantitative population parameter of interest describing the feature of the environment that the data user is investigating, such as a mean, median, or proportion,

L-2.1.2. It has a numerical value to which the parameter of interest will be compared, such as a regulatory or risk-based threshold or a similar parameter from another population (i.e., comparison to a reference site) or time (i.e., comparison to a prior time).

L-2.1.3. It has a relation that specifies precisely how the parameter will be compared to the numerical value, such as "is equal to" or "is greater than."

L-2.2. If the data user is interested in drawing inferences about only one population, the null and alternative hypotheses are stated in terms that relate the true value of the parameter to some fixed threshold value. A typical example of this one-sample problem in environmental studies is when the concentration of a contaminant is compared to a fixed regulatory limit or threshold value. For example, a data user may wish to determine whether the true mean concentration (μ) of the herbicide atrazine in groundwater at a hazardous waste site is greater than a fixed threshold value C, determined from a human or ecological risk assessment. If the decision maker wishes to "prove" that the contamination is less than C, it is initially assumed that the true (population) mean concentration is greater than or equal to C. This assumption is known as the null hypothesis and is denoted as H_0. If the data provide compelling evidence that the null hypothesis is false, then the null hypothesis is rejected and it would be concluded that the population mean concentration is less than C. The opposite conclusion is known as the alternative hypothesis and is denoted as H_A or H_1. For this example, the null and alternative hypotheses can be stated as follows:

$$H_0 : \mu \geq C, \quad H_A : \mu < C \ .$$

L-2.2.1. The null hypothesis (H_0) is the mean is greater than or equal to the threshold value *C*. The alternative hypothesis (H_A) is the opposite condition: the mean is less than the threshold value *C*.

L-2.2.2. If the decision maker wishes to demonstrate that the true mean is greater than the threshold value, the data must provide compelling evidence to reject this presumption, and the hypotheses can be stated as follows:

$$H_0 : \mu \leq C, \quad H_A : \mu > C \ .$$

L-2.2.3. Note that, thus far, two possible null hypotheses, $\mu \leq C$ and $\mu \geq C$, have been discussed. Depending upon the data quality objectives of the project, it is possible to legitimately assign either alternative to the null hypothesis. Because of this freedom or ambiguity, the most appropriate assignment must be determined from the project's data quality objectives.

L-2.2.4. Lastly, it should be noted that the null and alternative hypotheses for the examples presented above would be used for a one-sample, one-tailed statistical test. Typically, the sample mean of some set of measured concentrations would be statically compared to the threshold, *C*. The test is one-sample in nature because one data set (from one population) is used to calculate the test statistic, the sample mean. If, however, the statistical test entailed the use of two different data sets, in which each was potentially drawn from a separate population, it would be described as a two-sample test. The test is one-tailed in nature when the null hypothesis is an inequality. Although less common for environmental applications, the null and alternative hypotheses for the corresponding one-sample two-tailed test are as follows:

$$H_0 : \mu = C, \quad H_A : \mu \neq C \ \ (\text{i.e., } \mu > C \text{ or } \mu < C) \ .$$

L-2.2.5. The null hypothesis is that the population means is equal to *C* and the alternative hypothesis is that the population mean is either greater than or less than *C*.

L-2.3. If two populations are being compared, the null and alternative hypotheses are stated in terms that compare the true parameter value of one population to the corresponding true parameter value of the other population. A common example of this two-sample problem is when a potentially contaminated waste site is compared to a reference area using samples collected from the respective areas. In this situation, the hypotheses often are stated in terms of the difference between the two parameters; for example, the difference between the mean site concentration and the mean background concentration:

$$H_0 : \mu_{Site} - \mu_{Background} \leq 0, \quad H_A : \mu_{Site} - \mu_{Background} > 0 \ .$$

L-2.3.1. The hypothesis above would be used for a two-sample, one-tailed statistical test. As previously stated, the null and alternative hypotheses must be determined from project data quality objectives. Environmental regulations may specify particular null and alternative hypotheses. For example, the null hypothesis for a RCRA facility groundwater monitoring program is as follows: The concentration in down-gradient groundwater is less than or equal to the background concentration. When the null hypothesis is not specified by regulation, however, this determination should be made by carefully considering the consequences of making decision errors and taking the wrong actions. Selecting the null hypothesis is extremely important to the outcome of the decision process. The same set of sample data from a decision unit can lead to different decisions, depending on which possibility was selected as the null hypothesis.

L-2.3.2. Typically, hypothesis tests are established to prove a desired hypothesis. The condition or alternative that requires proof is selected as the alternative or research hypothesis. The alternative hypothesis is accepted (via burden of proof) when the null hypothesis is rejected (that is, disproved) based upon the weight of the evidence.

L-2.4. EPA 600/R-96/055, QA/G-4 recommends that the null hypothesis be defined as the true condition associated with the “more severe decision error”; that is, the more undesirable outcome if a wrong decision were made. For example, when the mean concentration of a contaminant is compared to a risk-based action level, *C*, the most severe decision error often consists of concluding $\mu < C$ when $\mu \geq C$ is the true condition. Therefore, as per EPA guidance, the null hypothesis is often $\mu \geq C$. In other words, it would typically be assumed that the site is “dirty” (H_0: $\mu \geq C$) until the weight of evidence demonstrates that the site is “clean” (H_A: $\mu < C$), the hypothesis that one wishes to demonstrate.

L-2.5. Rather than defining the null hypothesis based on the most severe condition, a second approach consists of defining the null hypothesis based on the least probable condition (or, equivalently, the alternative hypothesis based on the most probable condition). According to this approach, if a large amount of existing information suggests that one hypothesis is extremely likely, then this hypothesis would be defined as the alternative hypothesis. The advantage of this approach is that a large number of data may not be necessary to provide overwhelming evidence that the 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”

(H_A: $\mu \geq C$) and the null hypothesis as "the waste is not hazardous" (H_0: $\mu < C$). This approach generally will not result in the same null hypothesis as the approach EPA recommends. The most protective alternative for H_0 will not necessarily be the least probable alternative for H_0 (i.e., the most probable alternative for H_A).

L-2.6. Table L-1 summarizes common environmental decision rules and the corresponding hypotheses. The population parameter of interest (e.g., μ) in this table is denoted by the symbol Θ and the difference between two population parameters is denoted as $\Theta_1 - \Theta_2$, where Θ_1 represents the parameter of the first population (such as a constituent from a hazardous waste site) and Θ_2 represents the parameter of the second population (such as a constituent from background). The use of Θ is intended 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 as C, and the difference between two parameters is denoted as δ_0 (often the null hypothesis is defined such that $\delta_0 = 0$).

L-2.7. As previously discussed, hypothesis tests may be one-tailed or two-tailed, depending on the specified null and alternative hypotheses. The first, second, fourth, and fifth rows of Table L-1 are examples of one-tailed hypothesis tests. The third and sixth rows are examples of two-tailed tests. Most hypotheses connected with environmental monitoring are one-tailed because high pollutant levels can cause harm to humans or ecosystems, whereas lowered concentrations are of little, if any, concern.

THIS SPACE INTENTIONALLY LEFT BLANK

Table L-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: $\Theta_1 - \Theta_2 \leq \delta_0$ If $\delta_0 = 0$, H_0: $\Theta_1 - \Theta_2 \leq 0$ H_0: $\Theta_1 \leq \Theta_2$	H_A: $\Theta_1 - \Theta_2 > \delta_0$ If $\delta_0 = 0$, H_A: $\Theta_1 - \Theta_2 > 0$ H_A: $\Theta_1 > \Theta_2$
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: $\Theta_1 - \Theta_2 \geq \delta_0$ If $\delta_0 = 0$, H_0: $\Theta_1 - \Theta_2 \geq 0$ H_0: $\Theta_1 \geq \Theta_2$	H_A: $\Theta_1 - \Theta_2 < \delta_0$ If $\delta_0 = 0$, H_A: $\Theta_1 - \Theta_2 < 0$ H_A: $\Theta_1 < \Theta_2$
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: $\Theta_1 - \Theta_2 = \delta_0$ If $\delta_0 = 0$, H_0: $\Theta_1 - \Theta_2 = 0$ H_0: $\Theta_1 = \Theta_2$	H_A: $\Theta_1 - \Theta_2 \neq \delta_0$ If $\delta_0 = 0$, H_A: $\Theta_1 - \Theta_2 \neq 0$ H_A: $\Theta_1 \neq \Theta_2$

L-3. Decision Errors Associated with Hypothesis Tests. Table L-2 presents all of the possible scenarios that can result from a statistical hypothesis test. Two correct decisions and two incorrect decisions are possible. The probability of each event is presented in parenthesis.

Table L-2.
Conclusions Associated with Any Statistical Hypothesis Test

		True Hypothesis (Actual site conditions)	
		H_0 True	**H_a True**
Decision (Conclusion from sample data)	Do Not Reject H_0	Correct decision Confidence Level = $(1 - \alpha)100\%$	Incorrect decision False Acceptance of H_0 Type II error tolerance = β
	Reject H_0	Incorrect decision False Rejection of H_0 Type I error tolerance = α	Correct decision Power of test = $(1 - \beta)100\%$

L-3.1. The two incorrect answers for a hypothesis test are the following.

L-3.1.1. False rejection of H_0, or Type I error. The null hypothesis is rejected when the null hypothesis is true. The probability for a Type I error is defined as the level of significance. The maximum allowable probability for a Type I error is typically denoted by the symbol α. The level of confidence is defined as one minus the level of significance. Thus, the minimum level of confidence for a correct decision is $1 - \alpha$.

L-3.1.2. False acceptance or Type II error. The null hypothesis is accepted (more accurately, not rejected) when the null hypothesis is false. The maximum allowable probability for a Type II error is denoted by the symbol β. The power of the test is defined as one minus the Type II error probability. Therefore, the minimum power is $1 - \beta$.

L-3.2. A false rejection decision error occurs when it is concluded, from the observed data, that the null hypothesis is false when it is actually true. (This is sometimes called a "false positive.") A false acceptance decision error occurs when it is concluded that the null hypothesis is true when it is really false. (This is sometimes called a "false negative.") For example, suppose the null hypothesis states that the true value of the parameter of interest exceeds the action level. If the null hypothesis is actually correct and the sample data, by chance, contained an abnormally large proportion of low values, it would be concluded that the true value did not exceed the action level; therefore, a false rejection decision error would occur.

L-3.3. Three different equivalent approaches can be used to perform hypothesis tests: “The confidence interval,” “*p*-value,” and “critical value” approaches. Table L-3 illustrates the use of each of these three approaches for hypothesis testing.

Table L-3.
Relationship Between Hypothesis Tests and Confidence Intervals

Hypotheses	*p*-Value Approach Reject H_0 when	Critical Value Approach Reject H_0 when	Confidence Interval Approach Reject H_0 when
$H_0 : \Theta = C$ $H_A : \Theta \neq C$	$p < \alpha$	Test statistic less than or greater than critical values. Example: $t < t_{\alpha/2,n-1}$ or $t > t_{1-\alpha/2,n-1}$	Two-sided $1-\alpha$ confidence interval for Θ does not contain *C*
$H_0 : \Theta \geq C$ $H_A : \Theta < C$	$p < \alpha$	Test statistic less than “critical value. Example: $t < t_{\alpha,n-1}$.	One-sided $1-\alpha$ upper confidence interval limit for Θ is less than *C*: UCL < *C*
$H_0 : \Theta \leq C$ $H_A : \Theta > C$	$p < \alpha$	Test statistic exceeds “critical value.” Example: $t > t_{1-\alpha,n-1}$.	One-sided $1-\alpha$ lower confidence limit for Θ is greater than *C*: LCL > *C*

L-3.4. Table L-3 lists the possible null hypotheses for a one-sample statistical test. The objective is to determine if some population parameter of interest, Θ (the value of which is typically known) equals, is less than, or is greater than some fixed threshold value *C*. For the critical value approach for hypothesis testing, the decision to reject the null hypothesis is essentially determined by calculating some sample test statistic and comparing the value of the sample test statistic to a threshold or “critical value” for the sample statistic. If the sample statistic is greater than or less than the “critical value” (depending upon the null hypothesis selected), the null hypothesis is rejected.

L-3.5. Confidence intervals are directly related to hypothesis tests. Whenever a hypothesis test can be used to evaluate a parameter of interest (such as the mean, variance, median, etc.), a confidence interval also can be estimated and used to evaluate the same parameter. An equivalent approach consists of the following: Use the sample data to derive an estimate of the population parameter $\hat{\Theta}$, construct a confidence interval for Θ using the estimate $\hat{\Theta}$, and determine whether *C* falls within the confidence interval for Θ. If *C* does not fall within the confidence interval for Θ, then the null hypothesis is rejected. This is referred to as the “confidence interval approach for hypothesis testing.”

L-3.6. A third approach for hypothesis testing is referred to as the “*p*-value approach for hypothesis testing.” The “*p*-value” is the probability of obtaining the calculated sample statistic if the null hypothesis is true. If the *p*-value is sufficiently small, that is, if $p < \alpha$, where α is the Type I error tolerance, then the null hypothesis is rejected. All three approaches are illustrated below. This document predominately uses the critical value approach for hypothesis tests.

L-4. Illustration of Hypothesis Testing. To illustrate hypothesis testing, a one-population test to threshold value *C* is considered, with the following null and alternative hypotheses:

$$H_0 : \mu \geq C, \quad H_A : \mu < C \ .$$

Assume that the variable *X* is normally distributed with an unknown population mean μ but a known standard deviation σ. A single sample measurement *x* is compared to the threshold value, *C*, to determine whether or not to reject the null hypothesis, H_0: $\mu \geq C$. Because the standard deviation of the population (σ) typically would not be known for environmental applications, the example is not realistic, but serves only to illustrate the concept of hypothesis testing. Figure L-1 illustrates the decision errors for hypothesis testing.

L-4.1. Type I Error Tolerance and the Rejection of the Null Hypothesis. If the null hypothesis is true with $\mu = C$, a distribution of measured values of *X* would be obtained, as shown by the blue normal curve centered about $\mu = C$. The probability that a measurement, *x*, would be less than the critical value, X_α, is equal to α (refer to the region shaded in blue). The value of X_α depends upon the α value selected. The value of α is determined from the project's data quality objectives but is usually some acceptably small positive number (e.g., $\alpha = 0.01$ or 0.05). As the probability a measurement, *x*, will be less than X_α is acceptably small when $\mu = C$, the null hypothesis (H_0: $\mu \geq C$) is rejected when a measurement of $x < X_\alpha$ is obtained. (The null hypothesis is retained when $x > X_\alpha$.) The value α represents the tolerance for Type I error; that is, the maximum acceptable probability for rejecting H_0 when H_0 is actually true. When H_0 is $\mu \geq C$, the Type I error can be roughly described as the probability of concluding that a "dirty" site is "clean."

L-4.1.1. When *X* is normal with known standard deviation, σ, it is convenient to "standardize" the variable *X* using the linear transformation:

$$Z = \frac{X - \mu}{\sigma} \ .$$

L-4.1.2. The variable *Z* is a standard normal variable. If $x < X_\alpha$, it follows that

$$z = \frac{x - \mu}{\sigma} < Z_\alpha = \frac{X_\alpha - \mu}{\sigma} \ .$$

L-4.1.3. The quantity Z_α is the $\alpha100^{th}$ percentile of the standard normal distribution. Thus, if the null hypothesis $\mu = C$ is true and $x < X_\alpha$, then

$$z = \frac{x - C}{\sigma} < Z_\alpha \ .$$

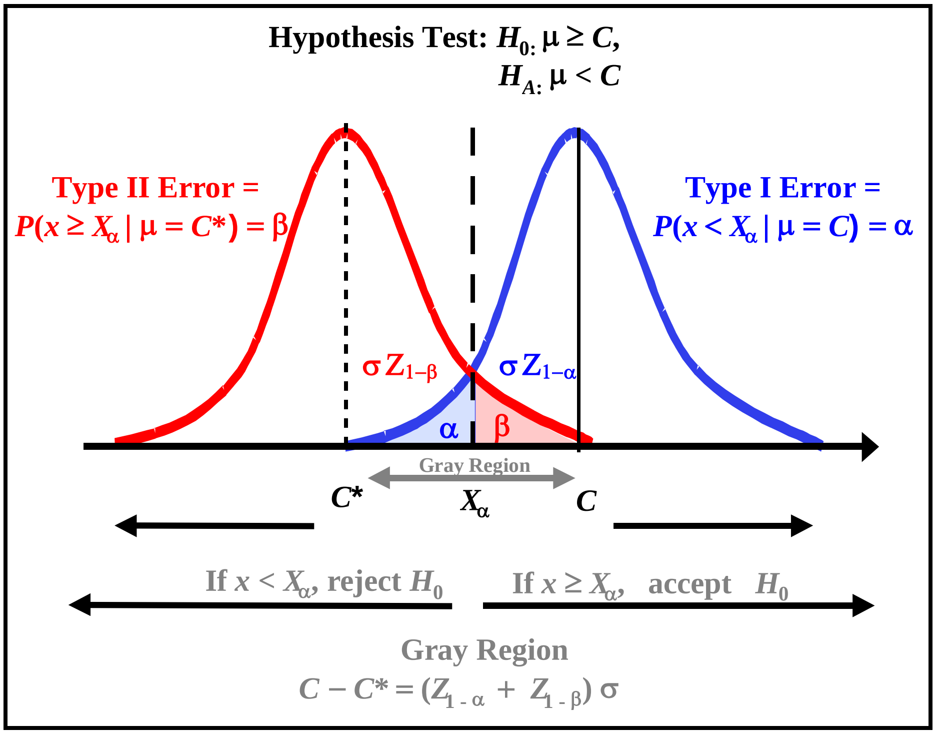

Hypothesis Test: $H_{0:}$ μ ≥ C,
$H_{A:}$ μ < C
Type II Error =
$P(x \geq X_\alpha \mid \mu = C^*) = \beta$
Type I Error =
$P(x < X_\alpha \mid \mu = C) = \alpha$
σ $Z_{1-\beta}$
σ $Z_{1-\alpha}$
α
β
Gray Region
C^*
X_α
C
If $x < X_\alpha$, reject H_0
If $x \geq X_\alpha$, accept H_0
Gray Region
$C - C^* = (Z_{1-\alpha} + Z_{1-\beta})\,\sigma$

Figure L-1. Decision Errors Associated with a Hypothesis Test.

L-4.1.4. Because H_0 is rejected when $x < X_\alpha$, it may be also be rejected when the test statistic $z < Z_\alpha$. In this context, the percentile Z_α is called the "critical value." If the sample statistic z is less than the "critical value" Z_α, it is often stated that the null hypothesis is rejected at the "α100% level of significance" or, equivalently, at the "(1 – α)100% level of confidence. " This is a convenient approach as the sample test statistic z can be calculated and compared to a desired percentile of the standard normal distribution (Z_α), which is readily available from a statistical table. The comparison of a sample statistic such as z to some percentile Z_α to determine whether or not to reject H_0 is referred to as the "critical value approach."

L-4.1.5. Statistical software provides an alternative to the critical value approach (for determining whether H_0 should be rejected), referred to as the "p value approach." For this particular example, given that a measure x from a normal distribution with known standard deviation (σ) is taken, the software also initially assumes that the null hypothesis is true (i.e., sets $\mu = C$), and calculates z. The calculated value is assumed to be equal to some percentile, $Z_{p,}$ of the standard normal distribution. Rather than reporting the statistic z and comparing it to the percentile Z_α, the software outputs the fraction of the normal probability distribution, p, that falls below the calculated value of z when $\mu = C$. This value is referred to as the "p value." The p value is the probability of obtaining a measured result of x (or a result different than the null hypothesis) when the null hypothesis is true ($\mu = C$). If p is sufficiently small relative to α (i.e., $p < \alpha$), the null hypothesis is rejected.

L-4.1.6. The third approach is referred to as the "confidence interval approach for hypothesis testing." It entails calculating a confidence interval for the population mean μ. In this situation, the best estimate of μ is the single measurement x. Because rejecting the null hypothesis requires

$$\frac{x - C}{\sigma} < Z_\alpha$$

and $Z_\alpha = -Z_{1-\alpha}$, it follows that the null hypothesis would be rejected if:

$$\text{UCL} = x + Z_{1-\alpha}\,\sigma < C \ .$$

L-4.1.7. The left side of the inequality is the one-sided $(1 - \alpha)100\%$ upper confidence limit for μ for a normal distribution with known standard deviation σ. Therefore, the null hypothesis is rejected if the UCL for μ is less than C. More information on confidence limits is contained in Appendix N.

L-4.1.8. The strategies discussed above are generally applicable for hypothesis tests, but the critical value approach is predominately used in this document.

L-4.2. Type II Error and Power. The discussion above focused on the criteria for rejecting the null hypothesis. The alternative hypothesis is discussed here. When the alternative hypothesis is true with $\mu = C^* < C$ (when the mean [μ] is equal to some value $C^* < C$), a normal distribution of measurements centered about $\mu = C^*$ will be obtained (refer to the red normal curve). When $\mu = C^*$, the probability $x > X_\alpha$ equals β (refer to the red shaded region). Because the null hypothesis is retained when $x > X_\alpha$, β is equal to the probability of retaining the null hypothesis (H_0: $\mu \geq C$) when the null hypothesis is false (i.e., when $\mu = C^* < C$). The value of β determined from project data quality objectives represents the maximum tolerance for Type II error; that is, the maximum tolerable probability for erroneously retaining the null hypotheses. In terms of an environmental investigation, the Type II error can be roughly described as the probability of concluding that a clean site is dirty. The power of the hypothesis test is defined as $1 - \beta$ and is equal to the probability of accepting the alternative hypothesis ($\mu = C^* < C$) when the alternative hypothesis is true (the probability of concluding that a clean site is clean).

L-4.2.1. Note that, to calculate the Type II error or the power of a test, the Type I error must first be specified. Also, note that, in this example, the Type II error tolerance and power is for some pre-specified value $C^* < C$. Paragraph L-5.2 illustrates how to calculate the power once α and C^* are specified for a normally distributed variable X with a known population standard deviation.

L-4.2.2. When the mean (μ) is equal to some value greater than C (when it falls somewhere to the right of C), the probability that the null hypothesis will be rejected is acceptably small, less than α. The probability that the null hypothesis will be retained will be greater than $1 - \alpha$. In terms of an environmental study, when $\mu > C$, the probability that a dirty site will be identified as dirty will be acceptably high. Similarly, when the mean (μ) is equal to some value less than C^*, the probability of retaining the null hypothesis (H_0: $\mu \geq C$) will be less than β. The probability of correctly rejecting the null hypothesis (and accepting H_A: $\mu < C$) will be greater than $1 - \beta$. When $\mu < C^*$, the probability that a clean site will be identified as clean will be acceptably high. However, when μ lies between C and C^*, the probability of making a correct decision will be low (the Type II error will be higher than β). This range of values, $C - C^*$, is called the “gray region” or the “minimum detectable difference.” Because reliable decisions cannot be made for differences smaller than $C - C^*$, the difference $C - C^*$ may be viewed as the “resolution” of the statistical design.

L-4.2.3. Statistical tests cannot control both types of error simultaneously. Generally, a hypothesis test is set up in a manner that committing false rejection (Type I) is considered the more serious error and is controlled by the test, and committing false acceptance (Type II) is considered not as serious an error and is not controlled by the test. The data user specifies the probability limit, α, by the data user’s tolerance for committing false rejection (Type I). Determining how large a risk the project team is willing to tolerate for Type I errors must be done before the fact, especially when the consequences of making such an error are very serious (Milton and Arnold, 1990). If the null hypothesis is not rejected after the test is performed, then the Type II error or the power (one minus the Type II error) is calculated. If the Type II error is not sufficiently small (or equivalently, the power is not sufficiently large), additional sampling would be considered. In general, increasing the sample size simultaneously reduces both Type I and Type II errors.

L-4.2.4. If the sample mean, $\bar{x}$, for a set of n measurements, rather than a single measurement, were compared to the threshold, C, to determine whether or not to reject the null hypothesis (H_0: $\mu \geq C$), then the minimum detectable difference would be given by:

$$C - C^* = \left(\sigma / \sqrt{n}\right)\left(Z_{1-\alpha} + Z_{1-\beta}\right) .$$

L-4.2.5. The number of random samples that must be collected can be solved from the above equation:

$$n = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2 \sigma^2}{\left(C - C^*\right)^2} \ .$$

L-4.2.6. Hence, the number of samples is dependent upon α, β, σ, and $C - C$*. The number of samples increases as the tolerance of Type I and Type II error, α and β, decreases (as $Z_{1-\alpha}$ and $Z_{1-\beta}$ increase as α and β decrease). The number of samples also increases as the variance (σ^2) increases and $C - C$* decreases. This is reasonable because the variance is a measure of the variability of the underlying environmental population and $C - C$* is a measure of the resolution of the statistical design. The number of samples increases as variability or heterogeneity of the underlying populations increases. As the probability of making a correct decision when the true mean lies in the gray area is low, the quantity $C - C$* essentially represents the smallest difference between the mean contaminant concentration and the threshold level that can be tolerated or that is deemed to be important for the overall statistical design. The sample size increases when smaller differences become significant for the statistical design.

L-5. Statistical Power Associated with Hypothesis Tests. As previously stated, the power of a statistical hypothesis test is defined as the likelihood that the null hypothesis is correctly rejected at a fixed level of significance, α, when the alternative hypothesis is truly correct. Power is related to Type II errors, or false rejection. The power of a statistical test is $1 - \beta$ where β is the probability of a false acceptance or Type II error. Therefore, as the power of a statistical test increases, the probability of a false acceptance decreases.

L-5.1. Introduction. To calculate the power of a statistical test, first determine the event that the test rejects the null hypothesis, H_0, in a form that does not contain any unknown parameters. There must be a predetermined level of significance, α, so there is a set criterion for rejecting the null hypothesis. The power is the calculated probability for rejecting the null hypothesis when the alternative hypothesis is assumed to be true. Unfortunately, the specific algorithm for calculating power is highly dependent upon the nature of the statistical test and power calculations are often complex. Paragraph L-5.2 presents directions for calculating the power for a hypothesis test of the form:

$$H_0 : \mu \le C, \quad H_A : \mu > C \ .$$

(Refer to Figure L-1.) The variable of interest is assumed to be normally distributed and the population standard deviation is known. The assumption that the population standard deviation (σ) is known severely limits the utility of the approach. However, it constitutes, perhaps, the simplest method to estimate power. In practice, an estimate of σ could be used to estimate the power if the uncertainty associated with the estimate was sufficiently small.

L-5.2. Example for Calculating the Power of a One-Tailed Test (from Mason et al., 1989). This procedure is strictly applicable only when the variable *X* is normally distributed with a known standard deviation. The procedure could potentially be used (to estimate the power) when the (population) standard deviation is not known and the sample is mean is calculated from a large number of samples (e.g., $n > 100$).

L-5.2.1. Suppose

$$H_0 : \mu \le 10, \quad H_A : \mu > 10 \ .$$

Assume a known standard deviation of $\sigma = 2$ for a normally distributed population. Let the Type I error tolerance for rejecting the null hypothesis $\alpha = 0.05$ and the sample size $n = 25$. Note that the threshold value $C = 10$. Let $C^* = 11$ in this example. Thus the "resolution" for the test, $C^* - C = 1$. Under the null hypothesis, the largest mean $\mu_0 = 10$. It follows that the power of the test is as follows:

$$1-\beta = P\left(\bar{x} > 10 \mid \mu = 11\right) = P\left\{\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha}\right\} = P\left\{\frac{\bar{x}-10}{2/\sqrt{25}} > 1.645\right\} = P\left\{\frac{\bar{x}-11}{2/\sqrt{25}} > 1.645 - \frac{11-10}{2/\sqrt{25}}\right\}$$
$$= P(Z > -0.855) = 1 - P(Z \le -0.855) = 0.804 \ .$$

$Z_{1-\alpha}$ is the $(1 - \alpha)100^{\text{th}}$ percentile of the standard normal distribution, which is provided in Table B-15 of Appendix B.

L-5.2.2. More generally, when comparing the sample mean (of a normally distributed variable with standard deviation σ) to some decision limit μ_0 using the null hypothesis, $H_0 : \mu \le \mu_0 = C$, the power at $\mu = \mu_1 = C^*$ is as follows:

$$1-\beta = P\left(\bar{x} > \mu_0 \mid \mu = \mu_1\right) = 1 - P\left\{Z \le \left(Z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)\right\} \ .$$

L-5.2.3. For this particular example, the experiment has a probability of 0.804 of correctly rejecting the null hypothesis when the true population mean is $\mu = 11$. If this power is not acceptably large, the sample size must be increased to maintain the same significance level. For example, a sample size $n = 50$ would produce the following power:

$$1-\beta = 1 - P\left\{Z \le \left(Z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)\right\} = 1 - P\left\{Z \le \left(1.645 - \frac{11-10}{2/\sqrt{50}}\right)\right\}$$
$$= 1 - P(Z \le -1.891) = 0.971 \ .$$

L-6. Tests for the Mean.

L-6.1. One-Sample t-test (Simple Random, Systematic Random, or Composite Sampling). 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 is parametric test that can be used to test hypotheses involving the mean (μ) of the population from which the sample was selected. The *t*-test is used when the population standard deviation is unknown but normality can be assumed.

L-6.1.1. Introduction.

L-6.1.1.1. The primary assumptions required for validity of the one-sample *t-test* are that the sample is random (data values are independent) and that the sample mean ($\bar{x}$) has an approximately normal distribution. Note that, according to the Central Limit Theorem, the sample mean will be approximately normally distributed for a large *n*. Unfortunately, the value of *n* that is sufficiently large enough to normalize the sample mean is seldom known. For environmental data, normality is not typically assumed for the sample mean unless *n* is very large (e.g., $n > 100$). Small sample sizes are common for environmental studies. As the sample mean is normal if *X* is normal, in practice, a data set consisting of *n* values of *X* is tested for normality and the *t*-test is used if the assumption of normality is not rejected.

L-6.1.1.2. Because the sample mean and standard deviation are very sensitive to outliers, the *t*-test should be preceded by a test for outliers (Appendix E). The *t*-test is also adversely affected by censored results. Directions for a one-sample *t*-test are presented in Paragraph L-6.1.2, followed by an example in Paragraph L-6.1.3.

L-6.1.2. Directions for a One-Sample t-test. The steps for a one-sample *t*-test are presented for Case 1: $H_0 : \mu \le C, H_A : \mu > C$; and Case 2: $H_0 : \mu \ge C, H_A : \mu < C$. The steps for Case 2 are given in braces {}. Let $x_1, x_2, \ldots, x_n$ represent the *n* data points from a normal distribution. These could be either *n* individual samples or *n* composite samples consisting of *k* aliquots each.

L-6.1.2.1. Verify that the data come from a normal distribution using tests presented in Appendices F and J, such as the Shapiro-Wilk test (Paragraph F-3.2) and a normal probability plot (Paragraph J-5.5).

L-6.1.2.2. Calculate the sample mean, $\bar{x}$, and the standard deviation, *s* (Appendix D).

L-6.1.2.3. Use Table B-23 of Appendix B to find the critical value, $t_{1-\alpha,\nu}$, such that 100(1 – α)% of the *t* distribution with $\nu = n-1$ degrees of freedom is below this value. For example, if α = 0.05 and *n* = 16, then *n* – 1 = 15 and $t_{0.95,15}$ = 1.753.

L-6.1.2.4. Calculate the test statistic *t* for the data set:

$$t = \frac{\overline{x} - C}{s/\sqrt{n}}.$$

L-6.1.2.5. Compare the calculated test statistic *t* with the critical value $t_{1-\alpha,\nu}$ (from Table B-23):

L-6.1.2.5.1. If $t > t_{1-\alpha,\nu}$ $\{t < -t_{1-\alpha,\nu}\}$, H_0 may be rejected. Go to L-6.1.2.7.

L-6.1.2.5.2. If $t \leq t_{1-\alpha,\nu}$ $\{t \geq -t_{1-\alpha,\nu}\}$, there is not enough evidence to reject H_0 and the false acceptance error rate should be verified. Go to L-6.1.2.6.

L-6.1.2.6. If H_0 is not rejected, calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates. The power of the test can be estimated using Paragraph L-5.2, assuming the true values for the mean and standard deviation are those obtained in the sample. A power curve of the test can be generated using software packages such as the Decision Error Feasibility Trial (DEFT) software (EPA QA/G-4D).

L-6.1.2.6.1. If only one false acceptance error rate (β) has been specified (at μ_1), it is possible to approximately 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. A derivation of the following formula is provided in Appendix A of EPA QA/G-4D.

L-6.1.2.6.2. 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 *p*100th percentile of the standard normal distribution (Table B-15, Appendix B).

L-6.1.2.6.3. Round *m* up to the next integer. If $m \leq n$, the false acceptance error rate has been satisfied. If $m > n$, the false acceptance error rate has not been satisfied.

L-6.1.2.7. Summary of results for one-sample *t*-test:

L-6.1.2.7.1. H_0 is rejected. One concludes $H_A : \mu > C$ $\{H_A : \mu < C\}$.

L-6.1.2.7.2. H_0 is not rejected and the false acceptance error rate is satisfied. One concludes $H_A : \mu \le C$ {$H_A : \mu \ge C$}; or

L-6.1.2.7.3. H_0 is not rejected but the false acceptance error rate is not satisfied. The null hypothesis must be retained but the conclusions are uncertain since the sample size is too small.

L-6.1.2.8. Report the results of the test, sample size, sample mean, standard deviation, and *t* and $t_{1-\alpha,\nu}$. Note that the calculations for the *t*-test are the same for both simple random or composite random sampling. The use of compositing usually results in a smaller value of *s* than simple random sampling.

L-6.1.3. Example of One-Sample t-Test for Simple and Systematic Random Samples with or without Compositing. Suppose total chromium in subsurface soil (below 5 feet from ground surface) at Site A is to be compared to a regulatory threshold of *C* = 2.0 mg/kg using the following test with 95% level of confidence:

$$H_0 : \mu \ge 2 , \quad H_A : \mu < 2 \ .$$

L-6.1.3.1. Table L-4 presents the data. All chromium concentrations were detected, so no proxy concentrations are needed to evaluate the data.

L-6.1.3.2. Verify that the data follow a normal distribution. The Shapiro-Wilk test for normality shows evidence that the data follow a normal distribution because the test's *p* value was 0.8489 and is > 0.05.

L-6.1.3.3. Calculate the mean and standard deviation: $\bar{x} = 4.619$ and $s = 0.8980$.

L-6.1.3.4. Because we want a 95% level of confidence, $\alpha = 0.05$. Also, because $n = 36$, $\nu = n - 1 = 36 - 1 = 35$.

L-6.1.3.5. Using Table B-23 of Appendix B and linear interpolation, the critical value is 1.6905.

$$t_{1-\alpha,\nu} = t_{0.95,35} = (1.697 + 1.684) / 2 = 1.6905 .$$

L-6.1.3.6. The test statistic is

$$t = \frac{\bar{x} - C}{s / \sqrt{n}} = \frac{4.619 - 2.0}{0.8980 / \sqrt{36}} = 17.50 .$$

L-6.1.3.7. Comparing the calculated test statistic, *t*, with the critical value, $t_{1-\alpha,\nu}$, we see that $t \ge -t_{1-\alpha,df}$ $(17.5 \ge -1.6905)$ and so we cannot reject H_0 and we must check that the false acceptance rate has been achieved.

Table L-4.
Example L-6.1.3 Data

Site A sample location	Top depth of sample (ft)	Bottom depth of sample (ft)	Chromium (total) concentration (mg/kg)		Site A sample location	Top depth of sample (ft)	Bottom depth of sample (ft)	Chromium (total) concentration (mg/kg)
EPC-SB01	9	10	2.95		EPC-SB07	9	10	5.1
EPC-SB01	14	15	5.17		EPC-SB07	14	15	4.94
EPC-SB01	19	20	4.8		EPC-SB07	19	20	4.76
EPC-SB02	9	10	4.53		EPC-SB08	9	10	4.62
EPC-SB02	14	15	4.01		EPC-SB08	14	15	4.72
EPC-SB02	19	20	5.91		EPC-SB08	19	20	4.73
EPC-SB03	9	10	3.96		EPC-SB09	9	10	3.21
EPC-SB03	14	15	4.81		EPC-SB09	14	15	4.14
EPC-SB03	19	20	5.27		EPC-SB09	19	20	4.85
EPC-SB04	9	10	5.99		EPC-SB10	9	10	4.25
EPC-SB04	14	15	4.6		EPC-SB10	14	15	5.09
EPC-SB04	19	20	5.51		EPC-SB10	19	20	3.68
EPC-SB05	9	10	4.72		EPC-SB11	9	10	5.12
EPC-SB05	14	15	3.56		EPC-SB11	14	15	6.6
EPC-SB05	19	20	4.22		EPC-SB11	19	20	6.19
EPC-SB06	9	10	3.91		EPC-SB12	9	10	3.15
EPC-SB06	14	15	5.81		EPC-SB12	14	15	4.11
EPC-SB06	19	20	4.48		EPC-SB12	19	20	2.8

L-6.1.3.8. Suppose the false acceptance rate is $\beta = 0.20$.

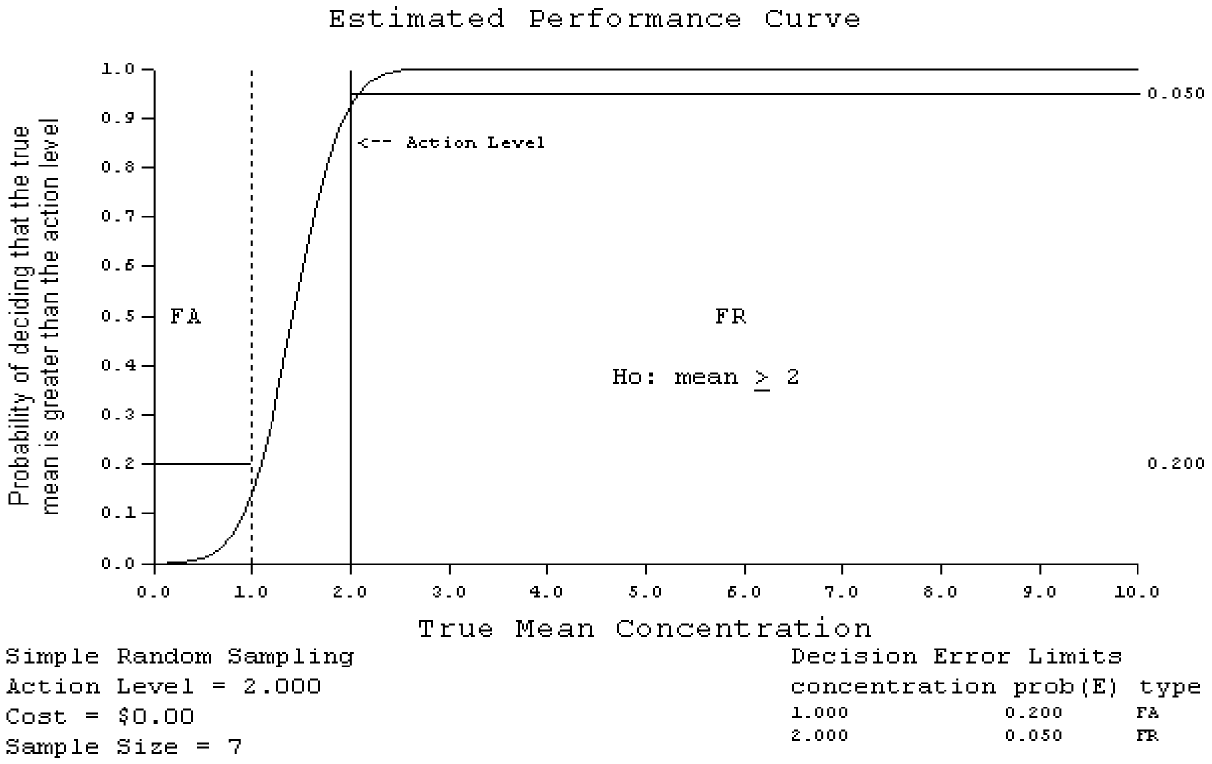
L-6.1.3.9. The power of this test is verified by assuming that the true values for the mean and standard deviation are those obtained in the sample. A power curve of the test was generated using DEFT software, as shown in the figure below. The probability of accepting the null hypothesis is plotted for a range of assumed true mean concentrations. For the regulatory threshold concentration of 2.0, a 95% (i.e., α = 0.05) chance of accepting the null hypothesis is requested. A 20% (β) probability of accepting the null hypothesis when the true concentration is μ_1 = 1.0 is also requested (80% power). A sample size of seven is suggested for this request. For the sample mean, this plot shows the probability of deciding that the true mean is higher than the regulatory threshold is nearly 100%, which means the test has strong power.

L-6.1.3.10. The sample size needed to achieve the false rejection rate of 0.20 when $\mu_1 = 1$ is:

$$m = \frac{s^2 (Z_{1-\alpha} + Z_{1-\beta})^2}{(\mu_1 - C)^2} + (0.5) Z_{1-\alpha}^2 = \frac{0.8980^2 (1.645 + 0.8417)^2}{(1-2)^2} + (0.5) 1.645^2 = 6.34.$$

Rounding up to the next integer, *m* = 7 (the reported value for “Sample Size” in Figure L-2).

L-6.1.3.11. Because more than seven samples have been collected (in fact, 36 samples have been collected), the false acceptance error rate has been satisfied. Therefore, we have evidence to suggest the true mean for chromium in Site A subsurface soil is greater than the regulatory threshold of 2.0 mg/kg on average.



```
Simple Random Sampling
Action Level = 2.000
Cost = $0.00
Sample Size = 7
```

Decision Error Limits

concentration	prob(E)	type
1.000	0.200	FA
2.000	0.050	FR

Figure L-2. Power Curve for the One-sample *t*-Test for Simple Random Sampling.

L-6.2. <u>One Sample t-Test for the Mean (Stratified Random Sampling)</u>. Directions for a one-sample *t*-test for a stratified random sample followed by an example are presented in Paragraphs L-6.2.1 and L-6.2.2, respectively.

L-6.2.1. Directions for a One-Sample t-Test for a Stratified Random Sample. The steps for a one-sample *t*-test are presented for: Case 1: $H_0 : \mu \le C, H_A : \mu > C$; and Case 2: $H_0 : \mu \ge C, H_A : \mu < C$. The steps for Case 2 are given in braces {}.

L-6.2.1.1. Let *h* = 1, 2, 3,…*L* represent the *L* strata and n_h represent the sample size of stratum *h*. The *i*th sample from stratum *h* is presented by $x_{h,i}$.

L-6.2.1.2. Verify that the data come from a normal distribution using tests presented in Appendices F and J, such as the Shapiro-Wilk test (Paragraph F-3.2) and a normal probability plot (Paragraph J-5.5).

L-6.2.1.3. Calculate the stratum weights w_h using the proportion of the volume in stratum *h*,

$$w_h = \frac{v_h}{\sum_{h=1}^{L} v_h}$$

where v_h is the surface area (or volume) of stratum *h* divided by the total surface area (or volume) over all strata.

L-6.2.1.4. For each stratum, calculate the sample stratum mean

$$\bar{x}_h = \frac{\sum_{i=1}^{n_h} x_{h,i}}{n_h}$$

and the sample stratum standard error

$$s_h^2 = \sum_{i-1}^{n_h} \frac{(x_{h,i} - \bar{x}_h)^2}{n_h - 1}.$$

L-6.2.1.5. Calculate overall mean and variance:

$$\bar{x}_{ST} = \sum_{h-1}^{L} w_h \bar{x}_h \,, \quad s_{ST}^2 = \sum_{h-1}^{L} w_h^2 \frac{s_h^2}{n_h}.$$

L-6.2.1.6. Calculate the degrees of freedom

$$v = \frac{\left(s_{ST}^2\right)^2}{\sum_{h=1}^{L} \frac{w_h^4 s_h^4}{n_h^2\left(n_h - 1\right)}} .$$

L-6.2.1.7. Use Table B-23 of Appendix B to find the critical value, $t_{1-\alpha,\nu}$, so that (1 – α)100% of the *t*-distribution with the above degrees of freedom (rounded to the next highest integer) is below $t_{1-\alpha,\nu}$.

L-6.2.1.8. Calculate the sample value (statistic):

$$t = \frac{\overline{x}_{ST} - C}{\sqrt{s_{ST}^2}} .$$

L-6.2.1.9. Compare the calculated test statistic, *t*, to the critical value $t_{1-\alpha,\nu}$,. If $t > t_{1-\alpha,\nu}$ {$t < -t_{1-\alpha,\nu}$,} H_0 may be rejected. If $t \leq t_{1-\alpha,\nu}$ {$\geq -t_{1-\alpha,\nu}$}, there is not enough evidence to reject H_0 and the false acceptance error rate should be verified.

L-6.2.1.10. If H_0 was not rejected, calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates. The results of the test could be:

L-6.2.1.10.1. H_0 was rejected so it seems that the true mean is less than *C* {greater than *C*}.

L-6.2.1.10.2. H_0 was not rejected and the false acceptance error rate was satisfied and it appears that the true mean is greater than *C* {less than *C*}; or,

L-6.2.1.10.3. H_0 was not rejected and the false acceptance error rate was not satisfied and it appears that the true mean is greater than *C* {less than *C*} but conclusions are uncertain since the sample size was too small.

L-6.2.1.10.4. If H_0 is not rejected, determine whether the power is adequate. Statistical software such as DEFT can be used for this purpose. DEFT uses the following approximation to calculate the number of samples required for each stratum to achieve a power of $1 - \beta$ at some desired value μ_1:

$$n_h' = \left[\sum_{h=1}^{L} w_h s_h\right] \times \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{\left(C - \mu_1\right)^2} \times w_h s_h,\ h = 1, \ldots, L .$$

The value n'_h is rounded up to a whole number. The power is adequate if the calculated sample size is less than or equal to the actual sample size for each stratum:

$$n'_h \leq n_h \text{ for } h = 1, \ldots, L.$$

L-6.2.2. Example of a One-Sample t-Test for a Stratified Random Sample. Suppose the total chromium in subsurface soil data used in the previous example (Paragraph L-6.2.1) came from a stratified sampling effort. Two strata were sampled, stratum A and stratum B, where stratum B makes up one-third of the area to be investigated. The objective is to compare the chromium concentration at Site A to a regulatory threshold of 2.0 mg/kg, based on a 95% level of confidence.

$$H_0 : \mu \geq 2, \quad H_A : \mu < 2 \ .$$

L-6.2.2.1. Table L-5 presents the data. All chromium concentrations were detected so no proxy concentrations are needed to evaluate the data.

$$L = 2 \qquad n_A = 24 \qquad n_B = 12 \qquad w_\mathrm{A} = 0.75 \qquad w_B = 0.25$$

L-6.2.2.2. Verify that the data follow a normal distribution for each stratum. The Shapiro-Wilk test was performed for each stratum and results indicated that the data for each follow a normal distribution because the tests' *p* values were greater than 0.05.

L-6.2.2.3. The mean and standard deviation of the data were calculated per stratum; $\alpha = 0.05$ because we want a 95% level of confidence:

$$\bar{x}_A = 4.674, \quad s_A = 1.027, \quad n_A = 24$$

$$\bar{x}_B = 4.508, \quad s_B = 0.5827, \quad n_B = 12$$

L-6.2.2.4. The overall mean and variance are:

$$\bar{x} = (0.75 \times 4.674) + (0.25 \times 4.508) = 4.633$$

$$s^2 = \left(0.75^2 \times \frac{1.027^2}{24} \right) + \left(0.25^2 \times \frac{0.5827^2}{12} \right) = 0.02472 + 0.001768 = 0.2649 \, .$$

L-6.2.2.5. The degrees of freedom are (rounded to the next highest integer):

$$\nu = \frac{\left(0.02649\right)^2}{\frac{0.75^4 \times 1.027^4}{24^2\left(24-1\right)} + \frac{0.25^4 \times 0.5827^4}{12^2\left(12-1\right)}} = 26.13 \approx 27 \ .$$

Table L-5.
Data for Example L-6.2.2

Stratum	Site A sample location	Top depth of sample (ft)	Bottom depth of sample (ft)	Chromium (total) concentration (mg/kg)	Stratum	Site A sample location	Top depth of sample (ft)	Bottom depth of sample (ft)	Chromium (total) concentration (mg/kg)
A	EPC-SB01	9	10	2.95	B	EPC-SB07	9	10	5.1
A	EPC-SB01	14	15	5.17	B	EPC-SB07	14	15	4.94
A	EPC-SB01	19	20	4.8	B	EPC-SB07	19	20	4.76
A	EPC-SB02	9	10	4.53	B	EPC-SB08	9	10	4.62
A	EPC-SB02	14	15	4.01	B	EPC-SB08	14	15	4.72
A	EPC-SB02	19	20	5.91	B	EPC-SB08	19	20	4.73
A	EPC-SB03	9	10	3.96	B	EPC-SB09	9	10	3.21
A	EPC-SB03	14	15	4.81	B	EPC-SB09	14	15	4.14
A	EPC-SB03	19	20	5.27	B	EPC-SB09	19	20	4.85
A	EPC-SB04	9	10	5.99	B	EPC-SB10	9	10	4.25
A	EPC-SB04	14	15	4.6	B	EPC-SB10	14	15	5.09
A	EPC-SB04	19	20	5.51	B	EPC-SB10	19	20	3.68
A	EPC-SB05	9	10	4.72	A	EPC-SB11	9	10	5.12
A	EPC-SB05	14	15	3.56	A	EPC-SB11	14	15	6.6
A	EPC-SB05	19	20	4.22	A	EPC-SB11	19	20	6.19
A	EPC-SB06	9	10	3.91	A	EPC-SB12	9	10	3.15
A	EPC-SB06	14	15	5.81	A	EPC-SB12	14	15	4.11
A	EPC-SB06	19	20	4.48	A	EPC-SB12	19	20	2.8

L-6.2.2.6. Table B-23 of Appendix B gives the critical value $t_{1-\alpha,\nu}$ = 1.703.

L-6.2.2.7. The test statistic is

$$t = \frac{\bar{x} - C}{s} = \frac{4.633 - 2.0}{\sqrt{0.02649}} \ .$$

L-6.2.2.8. Compare the calculated test statistic t with the critical value $t_{1-\alpha,\nu}$. Because $t \geq -t_{1-\alpha,\nu}$ $(16.18 \not\leq -1.703)$, we cannot reject H_0 and must check that the false acceptance rate has been achieved.

L-6.2.2.9. As in Paragraph L-6.1.3.9, a 20% (β) probability of accepting the null hypothesis when the true concentration is 1.0 is also requested (80% power). A power curve of the test was generated using DEFT software in Figure L-3 (by entering the sample standard deviation s_i and the weight w_i for each stratum). The required sample size for stratum A is equal to 5 and that for stratum B is equal to 2 (a total sample size of 7). The required power is achieved as actual the sample sizes for strata A and B are 24 and 12, respectively (a total of 36 samples).

L-6.3. The Chen Test. Environmental data such as concentration measurements are often confined to positive values and appear to follow a distribution with most of the data values relatively small or near zero, but with a few relatively large values. Underlying such data is a distribution that is not symmetrical (like a normal distribution) but is skewed to the right (like a lognormal distribution). Given a random sample of size *n* from a right-skewed distribution, the Chen test can be used to compare the mean (μ) of the distribution with a threshold level or regulatory value. This test assumes that the data arise from a right-skewed distribution and a random sample has been employed. Chen's test is a generalization of the *t*-test, with slightly more complicated calculations involving the sample mean, standard deviation, and skewness. Directions for conducting the Chen test are presented in Paragraph L-6.3.1, followed by an example in Paragraph L-6.3.2.

L-6.3.1. Directions for Conducting the Chen Test. Let $x_1, x_2, \ldots, x_n$ represent the *n* data points. Let *C* denote the threshold level of interest. The null hypothesis is $H_0 : \mu \le C$ and the alternative is $H_A : \mu > C$; the level of significance is α.

L-6.3.1.1. If, at most, 15% of the data points are below the detection limit and *C* is much larger than the DL, then replace values (< DL) with a proxy value (Appendix C).

L-6.3.1.2. Visually check the assumption of right-skewness by inspecting a histogram or frequency plot for the data.

L-6.3.1.3. Calculate the sample mean, $\bar{x}$, and the standard deviation, *s* (Appendix D).

L-6.3.1.4. Calculate the sample skewness

$$b = \frac{n\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{(n-1)(n-2)s^3}$$

the quantity

$$a = \frac{b}{6\sqrt{n}}$$

the statistic

$$t = \frac{(\bar{x} - C)}{s/\sqrt{n}}$$

and then compute:

$$z = t + a(1 + 2t^3) + 4a^2(t + 2t^3)\,.$$

The skewness, *b*, should be greater than 1 to confirm that the data are skewed to the right.

L-6.3.1.5. Use Table B-15 in Appendix B to find the critical value, $Z_{1-\alpha}$, such that $(1-\alpha)100\%$ of the standard normal distribution is below $Z_{1-\alpha}$, which is also the $p100^{th}$ percentile of the standard normal distribution. For example, if $\alpha = 0.05$ then $Z_{1-\alpha} = 1.645$.

L-6.3.1.6. Compare *z* with $Z_{1-\alpha}$:

L-6.3.1.6.1. If $z > Z_{1-\alpha}$, H_0 may be rejected and it appears that the true mean is greater than C.

L-6.3.1.6.2. If $z \le Z_{1-\alpha}$, there is not enough evidence to reject H_0 so it appears that the true mean is less than *C*.

L-6.3.2. Example of the Chen Test. Suppose surface soil samples (from 0 to 5 feet below ground surface) have been collected at Site B to evaluate arsenic concentrations on site against a regulatory threshold value of 5 mg/kg using a 90% level of confidence ($\alpha = 0.10$) and the following hypothesis test:

$$H_0 : \mu \le 5,\quad H_A : \mu > 5$$

Table L-6 presents the analytical results from samples collected at the site. All arsenic concentrations were detected so no proxy concentrations are needed to evaluate the data.

L-6.4. The Wilcoxon Signed Rank (One-Sample) Test. Given a random sample of size *n* (or composite sample size *n*, each composite consisting of *k* aliquots), the Wilcoxon signed rank test is a nonparametric test can be used to test hypotheses regarding the mean or median of the population from which the sample was selected. The mean is used as the parameter of interest in this Appendix, although the median could be used equivalently. The Wilcoxon signed rank test assumes that the data constitute a random sample from a symmetrical, continuous population. (Symmetrical means the underlying population frequency curve is symmetrical about its mean or median.) If the data are not symmetrical, it may be possible to transform them (using a transformation such as a log or square root transformation) so that this assumption is satisfied.

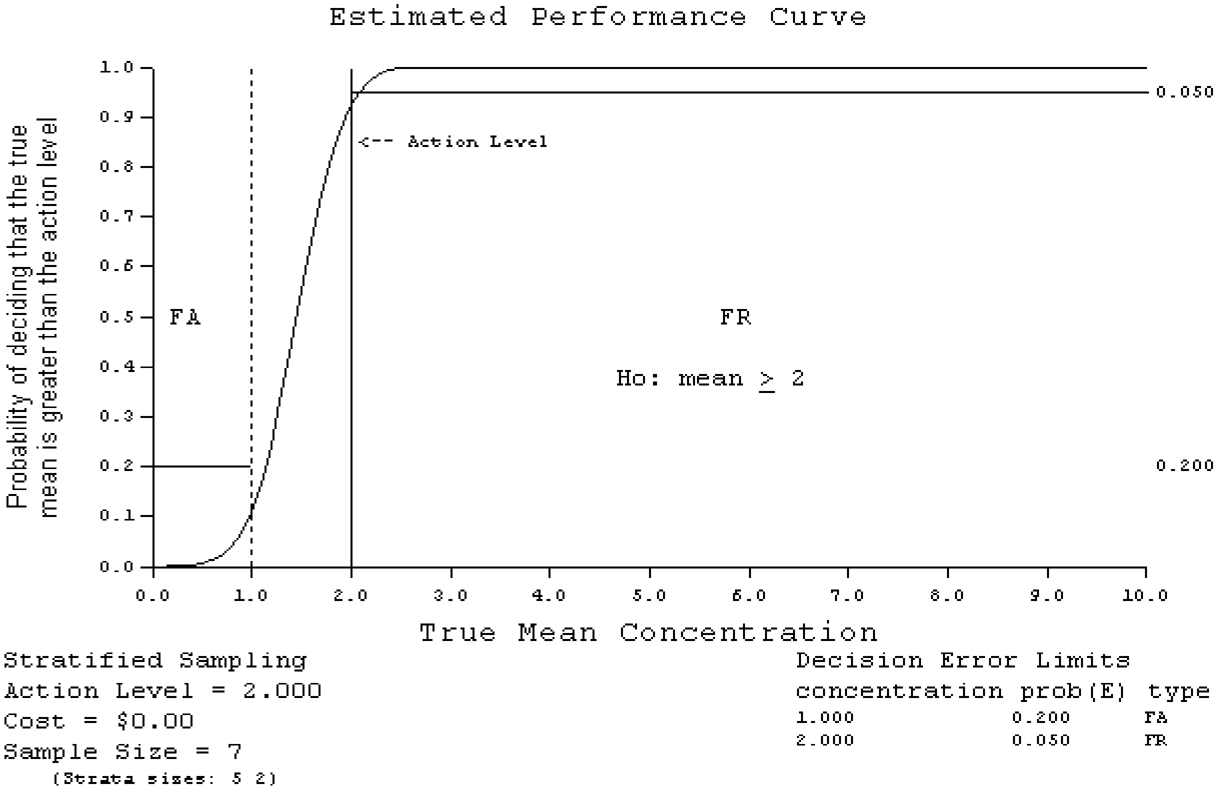


Figure L-3. Power Curve for the One-sample *t*-Test for Stratified Sampling.

L-6.4.1. Introduction. The Wilcoxon signed rank test is more robust to outliers. The *t*-test is not robust to outliers because the sample mean and standard deviation are strongly influenced by outliers. Although it is less powerful than the *t*-test when the data are normally distributed, it is usually more powerful when the data are not normally distributed. The Wilcoxon signed rank test is more likely than the *t*-test to identify differences for positively skewed distributions. In addition, compared to tests based on ranks, the *t*-test has difficulty accommodating censored values (values below the detection limit).

L-6.4.1.1. Directions for the Wilcoxon signed rank test for a simple random sample and a systematical simple random sample are given below in Paragraph L-6.4.2; Paragraph L-6.4.3 is an example for sample sizes smaller than 20.

L-6.4.1.2. 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 Paragraph L-6.4.4 followed by an example in Paragraph L-6.4.5.

L-6.4.1.3. Paragraph L-6.4.6 presents sample size calculations for the Wilcoxon signed rank test to achieve a certain power when the sample size is large. An example follows in Paragraph L-6.4.7.

Table L-6.
Analytical Results From Samples Collected at the Site in Example L-6.3.2

Site B sample location	Top depth of sample (ft)	Bottom depth of sample (ft)	Arsenic Concentration (mg/kg), x_i	$(x_i - \overline{x})^3$
EPC-BG01	1	2	4.84	–0.0024604
EPC-BG01	4	5	4.15	–0.5615156
EPC-BG02	1	2	4.53	–0.0881211
EPC-BG02	4	5	4.72	–0.0165814
EPC-BG03	1	2	4.76	–0.0099384
EPC-BG03	4	5	4.93	-9.112×10^{-5}
EPC-BG04	1	2	4.34	–0.2560479
EPC-BG04	4	5	4.51	–0.1005446
EPC-BG05	1	2	5.01	4.288×10^{-5}
EPC-BG05	4	5	3.83	–1.5011236
EPC-BG06	1	2	4.8	–0.0053594
EPC-BG06	4	5	4.07	–0.7412176
EPC-BG07	0.5	1	7.43	14.796346
EPC-BG07	2	2.5	4.6	–0.0527344
EPC-BG08	1	2	8.12	31.107274
EPC-BG08	4	5	4.96	-3.375×10^{-6}

L-6.4.2. Directions for the Wilcoxon Signed Rank Test for a Simple Random Sample and a Systematic Simple Random Sample. The following describes the steps for applying the Wilcoxon signed rank test for a sample size (*n*) less than 20 for: Case 1 ($H_0 : \mu \le C, H_A : \mu > C$); and Case 2 ($H_0 : \mu \ge C, H_A : \mu < C$). Modifications for Case 2 are given in braces {}.

L-6.4.2.1. Let $x_1, x_2, \ldots, x_n$ represent the *n* observations.

L-6.4.2.2. If possible, assign values to any measurements below the detection limit with procedures described in Appendix H.

L-6.4.2.3. Subtract *C* from each observation x_i to obtain the difference $d_i = x_i - C$. If any of the differences are zero, delete them and correspondingly reduce the sample size (*n*).

L-6.4.2.4. Assign ranks from 1 to *n* based on ordering the absolute differences $|d_i|$ (i.e., the 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 that otherwise would have been assigned to the tied observations (e.g., if two equal values occur after rank 5, then assign them each a rank of 6.5 = (6 + 7)/2).

L-6.4.2.5. Assign the sign for each observation to create the signed rank. The sign is positive if the deviation d_i is positive; the sign is negative if the deviation d_i is negative.

L-6.4.2.6. Calculate *R*, the sum of the ranks with a positive sign.

L-6.4.2.7. Use Table B-24 of Appendix B to find the critical value $w_{\alpha,n}$.

L-6.4.2.8. Compare the calculated test statistic, *R*, to the critical value.

L-6.4.2.8.1. If $R > n(n+1)/2 - w_{\alpha,n}$ $\{R < w_{\alpha,n}\}$, H_0 may be rejected.

L-6.4.2.8.2. If $R \le n(n+1)/2 - w_{\alpha,n}$ $\{R \ge w_{\alpha,n}\}$, there is not enough evidence to reject H_0.

L-6.4.2.9. The results of the test may be:

L-6.4.2.9.1. H_0 is rejected; $\mu > C$ $\{\mu < C\}$.

L-6.4.2.9.2. H_0 is not rejected $\mu \le C$ $\{\mu \ge C\}$.

L-6.4.3. Example of the Wilcoxon Signed Rank Test for Simple and Systematical Random Samples. Suppose $n = 14$ surface soil samples (from 0 to 5 feet below ground surface) were collected at Site B to evaluate cadmium concentrations on site against a regulatory threshold value of 0.75 using a 95% level of confidence ($\alpha = 0.05$) and the following hypothesis test.

$$H_0 : \mu \ge 0.75, \quad H_A : \mu < 0.75 .$$

L-6.4.3.1. Table L-7 presents the analytical results from samples collected at the site. Three of the cadmium concentrations were non-detects, so proxy concentrations are defined as the detection limit and are presented in parentheses.

L-6.4.3.2. Steps 1, 2, and 3 are contained in the three right-hand columns, in order.

L-6.4.3.3. Step 4: From the six cases where the sign of d_i is positive,

$$R = 13.5 + 13.5 + 12 + 11 + 2 + 20 = 62 \ .$$

L-6.4.3.4. Step 5: Table B-24 of Appendix B gives a critical value of $w_{0.05,14} = 26$.

Table L-7.
Analytical Results from Samples Collected at the Site in Example L-6.4.3

Site B sample location	Top depth of sample (ft)	Bottom depth of sample (ft)	Flag (ND = not detected)	Cadmium Concentration (mg/kg), x_i	$d_i = x_i - C$	Rank associated with $\lvert d_i \rvert$	Sign of d_i
EPC-BB01	1	2		1.6	0.85	13.5	+
EPC-BB01	4	5		1.6	0.85	13.5	+
EPC-BB02	1	2		1.55	0.8	12	+
EPC-BB02	4	5	ND	(0.242)	–0.508	9	–
EPC-BB03	1	2		0.624	–0.126	1	–
EPC-BB03	4	5		0.276	–0.474	7	–
EPC-BB04	1	2		1.5	0.75	11	+
EPC-BB04	4	5		0.301	–0.449	6	–
EPC-BB05	1	2		0.588	–0.162	3	–
EPC-BB05	4	5		0.264	–0.486	8	–
EPC-BB06	0.5	1		0.899	0.149	2	+
EPC-BB06	2	2.5		0.332	–0.418	4	–
EPC-BB07	1	2		1.42	0.67	10	+
EPC-BB07	4	5		0.326	–0.424	5	–

L-6.4.4. Directions for the Large Sample Approximation to the Wilcoxon Signed Rank Test. The following describes the steps for applying the large sample approximation of the Wilcoxon signed rank test for: Case 1 ($H_0 : \mu \le C, H_A : \mu > C$); and Case 2 ($H_0 : \mu \ge C, H_A : \mu < C$). Modifications for Case 2 are given in braces {}.

L-6.4.4.1. Let $x_1, x_2, \ldots, x_n$ represent the *n* data points where *n* is greater than or equal to 20. If possible, assign values to any measurements below the detection limit with procedures described in Appendix H.

L-6.4.4.2. Subtract *C* from each observation, x_i, to obtain the differences $d_i = x_i - C$. If any of the differences are zero delete them and correspondingly reduce the sample size (*n*).

L-6.4.4.3. Assign ranks from 1 to *n* based on ordering the absolute deviations $\lvert d_i \rvert$ (i.e., magnitude of differences ignoring the sign) from smallest to largest. Rank 1 is assigned to the smallest value, rank 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.

L-6.4.4.4. Assign the sign for each observation to create the signed rank. The sign is positive if the deviation, d_i, is positive; the sign is negative if the deviation, d_i, is negative.

L-6.4.4.5. Calculate the test statistic *R*, the sum of the ranks with a positive sign.

L-6.4.4.6. Calculate the critical value

$$w_p = n(n+1)/4 + Z_p \sqrt{n(n+1)(2n+1)/24}$$

where $p = 1 - \alpha \; \{p = \alpha\}$ and Z_p is the $100p^{th}$ percentile of the standard normal distribution (Table B-15 of Appendix B).

L-6.4.4.7. Compare the test statistic to the critical value. If $R > w_p \; \{R < w_p\}$, H_0 may be rejected. Otherwise, there is not enough evidence to reject H_0.

L-6.4.4.8. The results of the test may be:

L-6.4.4.8.1 H_0 is rejected; $\mu > C \; \{\mu < C\}$.

L-6.4.4.8.2 H_0 is not rejected; $\mu \leq C \; \{\mu \geq C\}$.

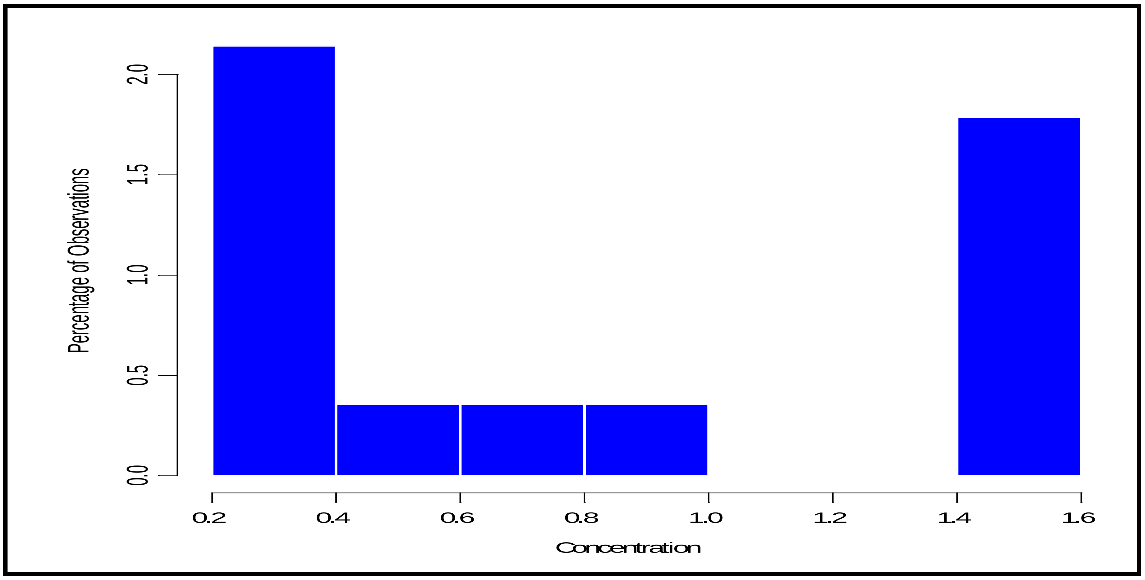


Figure L-4. Histogram Plot of Wilcoxon Signed Rank Test for Random Samples.

L-6.4.5. Example for the Large Sample Approximation to the Wilcoxon Signed Rank Test for Simple and Systematic Random Samples. Suppose additional surface soil samples (from 0 to 5 feet below ground surface) were collected at Site B to further delineate contamination. Additional samples were analyzed for cadmium and so the test performed earlier (see Paragraph L-6.4.3) for cadmium must be redone. The test was set up to compare cadmium concentrations on site to a regulatory threshold value of 0.75 using a 95% level of confidence ($\alpha = 0.05$) and the following hypothesis test.

$$H_0 : \mu \geq 0.75, \quad H_A : \mu < 0.75 \ .$$

L-6.4.5.1. Table L-8 presents all analytical results from samples collected from both sampling events. Non-detected cadmium concentrations were present in the data set; therefore, proxy concentrations are defined as the detection limit and are presented in parentheses.

L-6.4.5.2. Steps 1, 2, and 3 are contained in the three right-hand columns, in order.

L-6.4.5.3. Step 4: The test statistic, which is the sum of the ranks associated with the positive signs, is equal to

$$R = 21.5 + 21.5 + 20 + 19 + 3 + 17 + 18 + 15 + 1 + 16 = 152 \ .$$

L-6.4.5.4. Step 5: The critical value is

$$w_p = 22(22+1)/4 - 1.645\sqrt{22(22+1)(2 \times 22 + 1)/24} = 75.83$$

where $n = 22$ and by linear interpolation $Z_{0.05} = (-1.64 - 1.65)/2 = -1.645$.

L-6.4.5.5. Step 6: Comparing the test statistic to the critical value, $152 > 75.83, \left(R > w_p\right)$, so H_0 is not rejected.

L-6.4.5.6. Therefore, there is no evidence to suggest that the true mean for cadmium in Site B surface soil is less than the regulatory threshold of 0.75 mg/kg.

L-6.4.5.7. A histogram was created to check the symmetry of the data. The data appear symmetrical, as indicated in Figure L-3.

L-6.4.6. Directions for Calculating Sample Size for the Wilcoxon Signed Rank Test to Achieve a Specified Power. Noether (1987) discusses determining an adequate sample size based on a defined level of power to apply the Wilcoxon signed rank test for the following hypothesis test: Case 1 ($H_0 : \mu \le C$, $H_A : \mu > C$); and Case 2 ($H_0 : \mu \ge C$, $H_A : \mu < C$). Modifications for Case 2 are given in braces {}.

L-6.4.6.1. If the null hypothesis is not rejected, and the number of samples n' required to achieve some desired power $1-\beta$ could be calculated, the power would be adequate if $n \ge n'$. If $n \ge 20$ samples are collected, a conservative estimate of the sample size required for a power of $1-\beta$ is:

$$n' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{3\left(p' - 0.5\right)^2}$$

where Z_q is the q quantile of the standard normal distribution (from Table B-15), α is the significance level of the test, $1-\beta$ is the desired power for the test, and p' is the true probability that the average of any two independent observations

$$\frac{x_i + x_j}{2}$$

where $i \ne j$, exceeds {is less than} C.

Table L-8.
All Analytical Results from Samples Collected from Both Sampling Events

Site B sample location	Top depth of sample (ft)	Bottom depth of sample (ft)	Flag ND = not detected	Cadmium concentration (mg/kg), x_i	$d_i = x_i - C$	Rank associated with $\lvert d_i \rvert$	Sign of d_i
EPC-BB01	1	2		1.6	0.85	21.5	+
EPC-BB01	4	5		1.6	0.85	21.5	+
EPC-BB02	1	2		1.55	0.8	20	+
EPC-BB02	4	5	ND	(0.242)	–0.508	14	–
EPC-BB03	1	2		0.624	–0.126	2	–
EPC-BB03	4	5		0.276	–0.474	12	–
EPC-BB04	1	2		1.5	0.75	19	+
EPC-BB04	4	5		0.301	–0.449	10	–
EPC-BB05	1	2		0.588	–0.162	4	–
EPC-BB05	4	5		0.264	–0.486	13	–
EPC-BB06	0.5	1		0.899	0.149	3	+
EPC-BB06	2	2.5		0.332	–0.418	5	–
EPC-BB07	1	2		1.42	0.67	17	+
EPC-BB07	4	5		0.326	–0.424	8	–
EPC-BG08	1	2		1.48	0.73	18	+
EPC-BG08	4	5		0.302	–0.448	9	–
EPC-BG09	1	2		1.39	0.64	15	+

Site B sample location	Top depth of sample (ft)	Bottom depth of sample (ft)	Flag ND = not detected	Cadmium concentration (mg/kg), x_i	$d_i = x_i - C$	Rank associated with $\lvert d_i \rvert$	Sign of d_i
EPC-BG09	4	5		0.33	–0.42	6	–
EPC-BG10	0.5	1		0.812	0.062	1	+
EPC-BG10	2	2.5		0.287	–0.463	11	–
EPC-BG11	1	2		1.41	0.66	16	+
EPC-BG11	4	5		0.327	–0.423	7	–

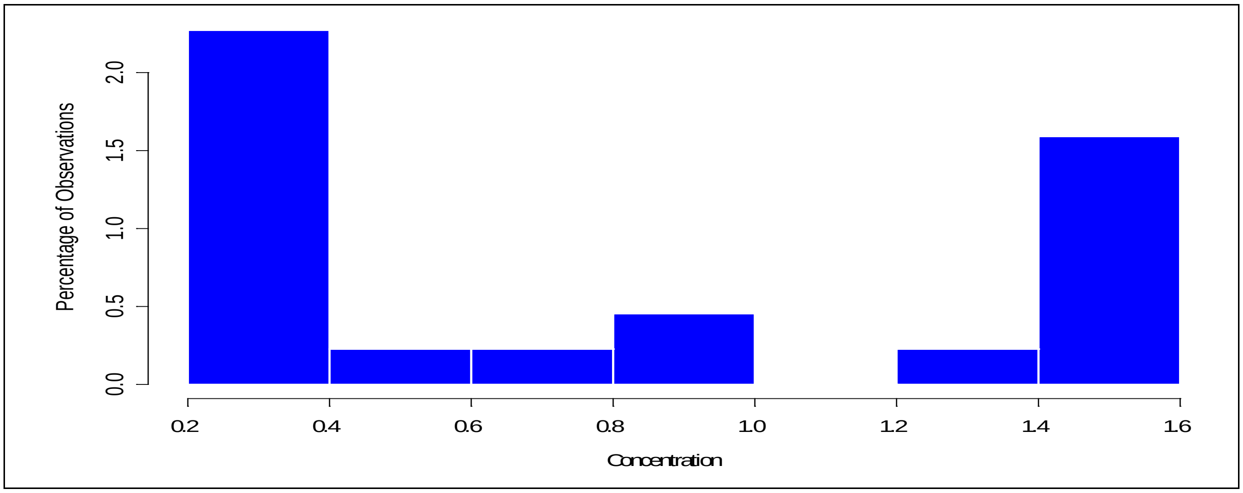


Figure L-5. Histogram Plot of Wilcoxon Signed Rank Test for Large Random Samples.

L-6.4.6.2. The equation for n' assumes that *n* is large enough for the test statistic *R* to be normally distributed (which is generally valid if the sample size exceeds 20). If the suggested sample size does not exceed 20, consult a statistician.

L-6.4.6.3. The value of p' can be determined from past information, a pilot sample, or chosen to represent a meaningful shift in the data (Noether, 1987). On the basis of what is considered to be a meaningful shift, one would assign p' equal to some probability greater than 0.5.

L-6.4.7. Example of Calculating Sample Size for the Wilcoxon Signed Rank Test to Achieve a Specified Power. Let us calculate the power for the hypothesis test performed in Paragraph L-6.4.5. In this example, *n* = 22 samples were collected to evaluate cadmium concentrations against a regulatory threshold value of 0.75 mg/kg at the 95% level of confidence ($\alpha = 0.05$) using the hypothesis test.

$$H_0 : \mu \geq 0.75,\ H_A : \mu < 0.75 \ .$$

The null hypothesis was not rejected. We wish to ensure that *n* is large enough to find a meaningful decrease in the mean with 80% probability (power).

L-6.4.7.1. The objective is to ensure that the sample size is large enough to find a meaningful decrease in the mean with 80% probability. Let us assume that seven samples had been collected for a prior "pilot" study. Table L-9 presents the analytical results from samples collected for the pilot study in the left-most column and along the top. The independent pair wise averages are calculated in the body of the table. Averages that fall below the regulatory threshold of 0.75 mg/kg are shaded.

Table L-9.
Analytical Results from Samples Collected for the Pilot Study and Independent Pair Wise Averages

Cadmium concentration (mg/kg)	1.220	0.301	0.624	0.276	0.588	0.264	0.332
1.220	—	0.761	0.922	0.748	0.904	0.742	0.776
0.301	—	—	0.463	0.289	0.445	0.283	0.317
0.624	—	—	—	0.450	0.606	0.444	0.478
0.276	—	—	—	—	0.432	0.270	0.304
0.588	—	—	—	—	—	0.426	0.460
0.264	—	—	—	—	—	—	0.298
0.332	—	—	—	—	—	—	—

L-6.4.7.2. Of the initial 7 results, 17 of the 21 independent averages are less than 0.75. The observed probability that the average of any two observed observations is less than *C* is 17/21 = 0.8095. Therefore, on the basis of this estimated (pilot study) probability, assume that it was determined that a power of 80% is required for $p' = 0.809$.

L-6.4.7.3. The required sample size to meet the power requirement is:

$$n' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{3\left(p' - \frac{1}{2}\right)^2} = \frac{(1.645 + 0.842)^2}{3(0.8095 - 0.5)^2} = 21.5 \ .$$

L-6.4.7.4. The required sample size is rounded up to 22. Because $n \geq n'$, the required power of 80% was achieved.

L-7. <u>Tests for a Median</u>. A population median ($\tilde{\mu}$) is another measure of the center of the population distribution. This population parameter is less sensitive than the sample mean to extreme values and non-detected results. Therefore, this parameter sometimes is used instead of the mean when the data contain a large number of non-detects or extreme values.

L-7.1. The Binomial Sign Test for the Median. Given a random sample of size *n* of continuous or discrete samples, the sign test may be used to test hypotheses regarding a population median for a distribution from which the data were drawn. The only assumption required for the sign test is that it be a random sample. The procedures are also robust to outliers, as long as they do not represent data errors. Directions for the sign test are given below in Paragraph L-7.2, followed by an example in Paragraph L-7.3.

L-7.2. Directions for the Sign Test for the Median. The following describes the steps for applying the sign test for a sample size (*n*).

Case 1 ($H_0 : \tilde{\mu}_x \leq C$ versus $H_A : \tilde{\mu}_x > C$); and

Case 2 ($H_0 : \tilde{\mu}_x \geq C$ versus $H_A : \tilde{\mu}_x < C$).

Modifications for Case 2 are given in braces {}. *C* is the hypothesized median or critical threshold value and $\tilde{\mu}_x$ is the median for the variable *X*. The level of significance is α.

L-7.2.1. Note that $\tilde{\mu}$ can also be defined as the median value for the variable *D*, where $D = X - C$ and so the hypotheses tests are written in terms of the difference.

Case 1 ($H_0 : \tilde{\mu}_D \leq 0$ versus $H_A : \tilde{\mu}_D > 0$); and

Case 2 ($H_0 : \tilde{\mu}_D \geq 0$ versus $H_A : \tilde{\mu}_D < 0$).

L-7.2.2 The hypotheses can also be written in terms of the probability of exceeding 0.

Case 1 ($H_0 : P(D \leq 0) \geq 0.5$ versus $H_A : P(D \leq 0) < 0.5$); and

Case 2 ($H_0 : P(D \geq 0) \geq 0.5$ versus $H_A : P(D \geq 0) < 0.5$).

Equivalently,

Case 1 ($H_0 : P(D > 0) \leq 0.5$ versus $H_A : P(D > 0) > 0.5$); and

Case 2 ($H_0 : P(D < 0) \leq 0.5$ versus $H_A : P(D < 0) > 0.5$).

This formulation suggests the use of the binomial distribution with *p* = 0.5 to test the null hypothesis.

L-7.2.3. Noether (1987) discusses determining an adequate sample size based on a defined level of power to apply the sign test for the median. Under the assumption that the test statistic (in this case the number of samples that exceed {are less than} *C*) is normally distributed, a conservative sample size, n', is calculated as:

$$n' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{4\left(p - \frac{1}{2}\right)^2}$$

where Z_q is the *q* quantile of the standard normal distribution (from Table B-15), α is the significance level of the test, $1-\beta$ is the desired power for the test, and p is the true probability that an observation exceeds {is less than} *C*. The value of p can be taken from past information, a pilot sample, or chosen to represent a meaningful shift in the data (Noether, 1987). The normality of the test statistic under the null hypothesis rests on the normal approximation to the binomial distribution. As discussed in Appendix E, this approximation works well when the sample size is at least 20 ($np \geq 10$, $p = 0.5$). If the suggested sample size does not exceed 20, consult a statistician.

L-7.2.4. Let $x_1, x_2, \ldots, x_n$ represent the *n* data points. Define a new variable $D = X - C$.

L-7.2.4.1. If possible, assign values to any measurements below the detection limit with procedures described in Appendix H. Subtract *C* from each observation, 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*).

L-7.2.4.2. Count the number of positive {negative} deviations (d_i) and denote this number by *y*.

L-7.2.4.3. The number of positive {negative} differences is described by a binomial distribution. In terms of the notation and terminology used in Appendix E, the number of data points is the number of "trials," *n*. Under the null hypothesis, the probability, *p*, of a positive {negative} difference (a success) is 0.5. The total number of positive {negative} differences, *y*, is the successful occurrence of an event *y* times out of *n*. Therefore, bin(*y*; *n*, *p* = 0.5) is the probability of *y* positive {negative} differences for a set of *n* trials, where the probability of a positive {negative} difference *p* = 0.5 (when H_0 is assumed to be true). The probability of obtaining less than or equal to *y* positive {negative} differences,

$$P(Y \leq y) = \sum_{i=0}^{y} bin(i, n, p)$$

is the value of the "cumulative binomial distribution." Table B-1 presents the probabilities of the cumulative binomial distribution for various values of *n, p,* and *k* where *k* = *y*.

L-7.2.4.4. If the probability of obtaining an equal or larger number of positive {negative} differences than the observed number *y* is small, that is, if $P(Y \geq y \mid n, p = 0.5) \leq \alpha$, then it is unlikely that the null hypothesis is true and the null hypothesis is rejected. Equivalently,

L-7.2.4.4.1. If $P(Y < y \mid n, p = 0.5) = P(Y \leq y - 1 \mid n, p = 0.5) \geq (1 - \alpha)$, H_0 may be rejected.

L-7.2.4.4.2. Otherwise, there is not enough evidence to reject H_0.

L-7.2.5. Use Table B-1 of Appendix B to find the probability value associated with $n,\ y - 1,$ and $p = 0.5$, which is the cumulative binomial distribution probability,

$$P(Y \leq y - 1 \mid n, p = 0.5)$$

to determine whether or not to reject the null hypothesis.

L-7.3. Example of the Sign Test for the Median. Suppose arsenic concentrations at a site are to be compared to a regulatory threshold value of 5 mg/kg using a 90% level of confidence ($\alpha = 0.10$). The median can be compared to this threshold using the following hypothesis test:

$$H_0 : \tilde{\mu} \leq 5,\ H_A : \tilde{\mu} > 5\ .$$

L-7.3.1. Suppose we wish to know the adequate sample size necessary to be 80% certain that we can detect a meaningful difference from the null hypothesis. The meaningful difference for this site is defined to be when the probability of exceeding the regulatory threshold is twice as likely as being below the threshold, $P(\tilde{\mu} > 5) = 2/3$. The required sample size is 41:

$$n' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{4\left(p - \frac{1}{2}\right)^2} = \frac{\left(Z_{0.9} + Z_{0.8}\right)^2}{4\left(\frac{2}{3} - \frac{1}{2}\right)^2} = \frac{\left(1.2816 + 0.8416\right)^2}{0.1111} = 40.6\ .$$

L-7.3.2. Consider the data presented in Paragraph L-6.3.2 for arsenic concentrations in surface soil samples (from 0 to 5 feet below ground surface) at Site B. Table L-10 presents the analytical results from samples collected at the site. All arsenic concentrations were detected, so no proxy concentrations are needed to evaluate the data.

L-7.3.3. The number of positive deviations (d_i), *y* = 3.

l-7.3.4. Using Table B-1 in Appendix B, we find $P(Y \le 2 \mid n = 16, p = 0.5) = 0.002090$

Table L-10.
Analytical Results From Samples Collected At The Site For Example L-7.3

Site B sample location	Top depth of sample (ft)	Bottom depth of sample (ft)	Arsenic concentration (mg/kg)	$d_i = x_i - C$	Sign of d_i
EPC–BG01	1	2	4.84	–0.16	–
EPC–BG01	4	5	4.15	–0.85	–
EPC–BG02	1	2	4.53	–0.47	–
EPC–BG02	4	5	4.72	–0.28	–
EPC–BG03	1	2	4.76	–0.24	–
EPC–BG03	4	5	4.93	–0.07	–
EPC–BG04	1	2	4.34	–0.66	–
EPC–BG04	4	5	4.51	–0.49	–
EPC–BG05	1	2	5.01	0.01	+
EPC–BG05	4	5	3.83	–1.17	–
EPC–BG06	1	2	4.8	–0.2	–
EPC–BG06	4	5	4.07	–0.93	–
EPC–BG07	0.5	1	7.43	2.43	+
EPC–BG07	2	2.5	4.6	–0.4	–
EPC–BG08	1	2	8.12	3.12	+
EPC–BG08	4	5	4.96	–0.04	–

L-7.3.5. As 0.002090 < 0.9, H_0 may not be rejected. Therefore, it appears that the true median for arsenic is less than the regulatory threshold of 5 mg/kg. However, to achieve 80% power and satisfy the sample size requirement calculated earlier, an additional 25 randomly selected samples would be needed to increase the total sample size to 41.

L-8. Test for a Proportion or Percentile.

L-8.1. The One-Sample Proportion Test. Given a random sample of size *n*, the nonparametric, 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. The only assumption required for the one-sample proportion test is that it be 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 in sample collection.

L-8.1.1. The null and alternative hypotheses for this test can be stated as:

$$H_0 : X_{Po} \le C\,,\quad H_A : X_{Po} > C$$

where X_{Po} is the P_0 quantile of the variable *X*; that is,

$$P(X \leq X_{Po}) = P_0 \ .$$

L-8.1.2 If *P* is the “true” proportion of *X* that is less than or equal to *C* = X_P, then

$$P(X \leq C) = P \ .$$

L-8.1.3. The hypothesis statement can be written as:

$$H_0 : P_0 \leq P \, , \ \ H_A : P_0 > P \, .$$

L-8.1.4. Equivalently,

$$H_0 : P \geq P_0 \, , \ \ H_A : P < P_0 \ .$$

(Note that *P*, the true portion of the population less than *C*, should not be confused with the probability density function *P*(*X*) for the variable *X* discussed in Appendix E.)

L-8.1.5. Because the only assumption is that it be 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. This test is recommended when fewer than 50% of the results are detected. The test may be used as long as the proportion of non-detects is smaller than the proportion, p_0, of interest, and *n* must be relatively large for the test to be reliable.

L-8.1.6. Directions for the one-sample proportion test for a simple random sample and a systematic random sample are given below in Paragraph L-8.2, followed by an example presented in Paragraph L-8.3.

L-8.2. <u>Directions for a Simple Random Sample and a Systematic Random Sample</u>. Directions to apply the one-sample proportion test for Case 1 and Case 2: Case 1 ($H_0 : P \leq P_0 \, , H_A : P > P_0$); and Case 2 ($H_0 : P \geq P_0 \, , H_A : P < P_0$), which are given in braces { }.

L-8.2.1. Given a random sample $x_1, x_2, \ldots, x_n$ of measurements from the population, let *P* denote the proportion of *X*'s that do not exceed *C*. This true proportion can be estimated from the sample data by dividing the number (*k*) of sample points that are less than or equal to *C* by the sample size (*n*).

$$P \approx p = k/n \ .$$

L-8.2.2. Compute *np*, and *n*(1 – *p*). If both *np* and *n*(1 – *p*) are greater than or equal to 5, proceed.

L-8.2.3. Otherwise, consult a statistician as analysis may be complex. Calculate:

$$z = \frac{p - P_0}{\sqrt{P_0(1 - P_0)/n}}.$$

L-8.2.4. Use Table B-15 of Appendix B to find the critical value, $Z_{1-\alpha}$, such that $(1-\alpha)100\%$ of the normal distribution is below $Z_{1-\alpha}$. For example, if $\alpha = 0.05$ then $Z_{1-\alpha} = 1.645$.

L-8.2.4.1. If $z > Z_{1-\alpha}\{z < -Z_{1-\alpha}\}$, H_0 may be rejected.

L-8.2.4.2. If $z \le Z_{1-\alpha}\{z \ge -Z_{1-\alpha}\}$, there is not enough evidence to reject H_0. Therefore, the false acceptance error rate must be verified.

L-8.2.5. To calculate the power of the test, choose a proportion, P_1, that would constitute a meaningful difference from P_0, and use a statistical software package to generate the power curve of the test.

L-8.2.6. If only one false acceptance error rate (*β*) has been specified (at P_1), it is possible to calculate the sample size that achieves the DQOs. To do this, calculate:

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

L-8.2.7. If $m \le n$, the false acceptance error rate has been satisfied. Otherwise, the false acceptance error rate has not been satisfied. It is usually more helpful to do this calculation before sampling, as all of the parameter values needed for the calculation are available before the sampling begins.

L-8.2.8. The results of the test could be:

L-8.2.8.1. H_0 is rejected, conclude that $P > P_0$ $\{P < P_0\}$.

L-8.2.8.2. H_0 is not rejected, the false acceptance error rate was satisfied, and conclude that $P \le P_0$ $\{P \ge P_0\}$.

L-8.2.8.3. H_0 is not rejected, the false acceptance error rate was not satisfied, and the conclusion that $P \le P_0$ $\{P \ge P_0\}$ is uncertain because the sample size was too small.

L-8.2.9. Example of the One-Sample Test for Proportions of Simple and Systematic Random Samples. Groundwater concentrations of gasoline at a site are compared to a regulatory threshold C = 35 micrograms per liter (μg/L). Suppose this site has only 13 detections out of 90 groundwater samples collected to date. Because more than 50% of the data are censored, the test of proportions is more appropriate than a *t*-test or Wilcoxon signed rank test. The test of proportions can be used to determine if more than 95% of the concentrations are less than the regulatory threshold at the 90% level of confidence. The null and alternative hypotheses are as follows:

$$H_0 : X_{0.95} \ge 35\ \mu\text{g/L},\ H_A : X_{0.95} < 35\ \mu\text{g/L}\ .$$

L-8.2.9.1. Equivalently,

$$H_0 : P \le 0.95,\quad H_A : P > 0.95\ .$$

(This is Case 1 in Paragraph L-8.2.) Suppose 11 of the detected concentrations exceed this regulatory threshold; therefore, the proportion of samples with detected concentrations below the threshold is $p = (90 - 11)/90 = 0.8778$.

L-8.2.9.2. Determine whether $np \ge 5$ and $n(1 - p) \ge 5$:

$$np = 90 \times 0.8778 = 79$$

$$n(1 - p) = 90 \times (1 - 0.08778) = 11\ .$$

L-8.2.9.3. Because $np \ge 5$ and $n(1 - p) \ge 5$, the test of proportions can be used. In this example, P_0 = 0.95 and $1 - \alpha = 0.90$.

$$z = \frac{p - P_0}{\sqrt{P_0(1 - P_0)/n}} = \frac{0.8778 - 0.95}{\sqrt{0.95(1 - 0.95)/90}} = -3.143\,.$$

L-8.2.9.4. Using Table B-15 of Appendix B, we find the critical value $Z_{0.90}$=1.282.

L-8.2.9.5. Compare the calculated value z with the critical value. The null hypothesis is rejected if $z > Z_{0..90}$. As $-3.143 \le 1.282$ $(z \le Z_{0.90})$, there is not enough evidence to reject H_0. Therefore, the false acceptance error rate has to be verified through a power curve or sample size calculation. Suppose a false acceptance error rate was specified at

$P_1 = 0.99$ ($\beta = 0.20$); it is possible to calculate the sample size that achieves this error rate using the following equation:

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

$$= \left[\frac{1.282\sqrt{0.95(1-0.95)} + 0.8417\sqrt{0.99(1-0.99)}}{0.99-0.95} \right]^2 = 82.43 \approx 83.$$

L-8.2.9.6. Because $83 \le 90$ $(m \le n)$, the false acceptance error rate has been satisfied. Therefore, H_0 was not rejected and the false acceptance error rate was satisfied. There is at least 90% confidence that the proportion of gasoline concentrations below the regulatory threshold is less than 0.95 (i.e., $P \le 0.95$, or, equivalently, $X_{0.95} \ge 35$).

APPENDIX M

Hypothesis Testing—Two-Population and General Cases

M-1. Introduction. A two-sample test is used when a data user is interested in making inferences about two independent populations, comparing some parameter from one population to the corresponding parameter from a second population. For example, a common environmental application entails comparing the population mean or median of the study area data set to the population mean or median of the background data set. EPA 600/R-96/084, QA/G-9 contains additional examples of the basic statistical tests presented here. Lehmann (1975) is a good resource for nonparametric tests. Montgomery (1997) contains a fuller treatment of two-sample *t*-tests, matched pairs *t*-tests, ANOVA, and multiple comparison tests.

M-2. Comparing Two Means. Two-sample tests do not require equal sample sizes, though equal sample sizes are recommended. The accuracy of estimating summary statistics from each sample is based on the number of samples available; data sets with many samples can provide more accurate estimates of the mean and standard deviation than those with only a few. When sample sizes are not equal, it may mean that one population is not defined as well as the other. If sample sizes are grossly unequal, the result of the two-sample test may produce an incorrect conclusion.

M-2.1. Student's Two-Sample t-Test. Student's two-sample *t*-test is a parametric statistical test that can be used to compare two population means based on the independent random samples $x_1, x_2,..., x_m$ from the first population, and samples $y_1, y_2,..., y_n$ from the second population. This test assumes the variances of the two populations are approximately equal. This supposition can be verified using an *F*-test or Levene’s test (Appendix N, Paragraph N-4). However, the *F*-test is not recommended because it is not robust to deviations from normality. A positively skewed distribution tends to give rise to higher values of *F* and false rejection of the null hypothesis that the variances of two distributions are equal. If the two variances are not equal, the Satterthwaite’s *t*-test is recommended (See Paragraph M-2.1.2 for directions and Paragraph M-2.1.3 for an example).

M-2.1.1. Introduction. The principal assumption required for the two-sample *t*-test is that a random sample of size *m* ($x_1, x_2,..., x_m$) is drawn from population 1, and an independent random sample of size *n* ($y_1, y_2,..., y_n$) is drawn from population 2. The second assumption required for the two-sample *t*-test is that the sample means, $\bar{x}$ (sample 1) and $\bar{y}$ (sample 2), are approximately normally distributed (if *X* and *Y* are normal, the sample means $\bar{x}$ and $\bar{y}$ will be also be normally distributed).

M-2.1.1.1. The two-sample *t*-test is commonly used to compare site contaminant concentrations to background concentrations:

$$H_0 : \mu_S - \mu_B \leq \delta_0,\ H_A : \mu_S - \mu_B > \delta_0\ .$$

The "true" mean site concentration and "true" mean background concentrations are denoted by μ_S and μ_B, respectively. When the above null hypothesis is selected, often δ_0 = 0 and α = 0.2 or 0.1. For this situation, the value of α tends to be somewhat higher than that used for other statistical applications (e.g., where α may be 0.05 or 0.01). This occurs to avoid a large Type II error (in this case, concluding the site is "clean" when it is "dirty" relative to background). As α decreases, the value of $\bar{x} > \bar{y}$ required to reject $H_0 : \mu_X \leq \mu_Y$ increases. The following null and alternative hypotheses are also frequently used:

$$H_0 : \mu_S - \mu_B \geq \delta_0,\ H_A : \mu_S - \mu_B < \delta_0\ .$$

M-2.1.1.2. In this situation, a common value for α is 0.05. However, the value for δ_0 depends greatly on the project. To reject H_0, that is, to demonstrate that the site is "clean" relative to background, the site mean must be significantly less than the background plus δ_0 (e.g., $\bar{x} << \bar{y} + \delta_0$). When there is actually no difference between the site and background populations (i.e., $\mu_S = \mu_B$), rejecting the null hypothesis in favor of the alternative hypothesis (i.e., the site is "clean" relative to background), becomes less probable as the selected value of δ_0 decreases. In general, a small value of δ_0 is undesirable from a cost perspective as a larger than budgeted number of samples may be required to determine if the means differ by δ_0. However, an extremely large value of δ_0 is undesirable from an environmental risk perspective as H_0 may be rejected even when the site mean is much larger than the background mean. Occasionally, δ_0 is equal to one or two standard deviations of the background data set. The selection of an appropriate value of δ_0 is a critical component of the DQO process during project planning; the value should be established only after input is obtained from all users and stake holders.

M-2.1.2. Directions to Apply the Two-sample t-test for Differences Between the Population Means. Steps to apply the two-sample *t*-test for differences between the population means for Case 1 and Case 2 are as follows: Case 1: $H_o : \mu_X - \mu_Y \leq \delta_o$, $H_A : \mu_x - \mu_y \geq \delta_0$; and Case 2: $H_0 : \mu_x - \mu_y \geq \delta_0$, $H_A : \mu_x - \mu_y \leq \delta_0$, which is given in braces { }.

M-2.1.2.1. Verify that both data sets are normal, using procedures in Appendices F and J, such as the Shapiro-Wilk test (Paragraphs F-3.2 and F-3.3) and a normal probability plot (Paragraphs J-5.5 and J-5.6).

M-2.1.2.2. Calculate the sample mean, $\bar{x}$, and the sample variance, s_X^2 (Appendix D), for the first data set (containing *m* points) and compute the sample mean, $\bar{y}$, and the sample variance, s_Y^2, for the second data set (containing *n* points).

M-2.1.2.3. 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 (presented below). Otherwise, compute the pooled standard deviation:

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

M-2.1.2.4. Calculate

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{s_E \sqrt{1/n + 1/m}}\ .$$

M-2.1.2.5. Use Table B-23 of Appendix B to find the critical value, $t_{1-\alpha,m+n-2}$, such that $(1-\alpha)100\,\%$ of the *t*-distribution with (m + n – 2) degrees of freedom is below $t_{1-\alpha,m+n-2}$.

M-2.1.2.5.1. If $t > t_{1-\alpha,m+n-2}\{t < -t_{1-\alpha,m+n-2}\}$, reject H_0. Go to step M-2.1.2.7.

M-2.1.2.5.2. If $t \le t_{1-\alpha,m+n-2}\{t \ge -t_{1-\alpha,m+n-2}\}$, there is not enough evidence to reject H_0 . Therefore, the false acceptance error rate will need to be verified. Go to M-2.2.6.

M-2.1.2.6. 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. If only one false acceptance 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.

M-2.1.2.7. Calculate:

$$m^* = n^* = \frac{2 s_E^2 \left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\left(\delta_1 - \delta_0\right)^2} + (0.25)\, z_{1-\alpha}^2\ .$$

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

M-2.1.2.8. The results of the test could be:

M-2.1.2.8.1. H_0 is rejected; $\mu_x - \mu_y > \delta_0\{\mu_x - \mu_y < \delta_0\}$.

M-2.1.2.8.2. H_0 is not rejected and the false acceptance error rate is satisfied; $\mu_x - \mu_y \le \delta_0 \{\mu_x - \mu_y \ge \delta_0\}$.

M-2.1.2.8.3. H_0 is not rejected and the false acceptance error rate was not satisfied; $\mu_x - \mu_y \le \delta_0 \{\mu_x - \mu_y \ge \delta_0\}$, but this conclusion is uncertain because the sample size was too small.

M-2.1.3. Example of the Student's Two-Sample t-Test (Equal Variances) for Simple and Systematic Random Samples. Consider the case where nickel (Ni) surface soil concentrations are compared between Site A and Background using the test:

$$H_o : \mu_X - \mu_Y \le \delta_o,\;\; H_A : \mu_x - \mu_y > \delta_0\;.$$

Let *X* refer to the site Ni concentrations and *Y* to the background Ni concentrations.
Let $\delta_0 = 0$.

M-2.1.3.1. The following Ni concentrations are obtained for the site soil ($m = 6$): 2.665, 3.610, 5.470, 7.150, 8.340, and 7.960 mg/kg.

M-2.1.3.2. The following Ni concentrations are obtained for the background soil ($n = 10$): 5.140, 7.460, 5.990, 3.360, 3.190, 2.870, 5.950, 1.720, 4.770, and 5.605 mg/kg.

M-2.1.3.3. In this example, the Shapiro-Wilk test was used to test the assumption of normality and an *F*-test was used to test the assumption of equal variances. Because the data have equal variances at a significance level of 0.05, the Student’s two-sample *t*-test is more appropriate.

Data	**Sample Mean**	**Sample Variance**	**Sample Size**
Site (*X*)	5.87	5.53	6
Background (*Y*)	4.61	3.12	10

M-2.1.3.4. Using methods presented above in Paragraph M-2.1, 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 (Paragraph M-2.2). Otherwise, compute the pooled standard deviation:

$$s_E = \sqrt{\frac{(m-1)s_X^2 + (n-1)s_Y^2}{(m-1) + (n-1)}} = \sqrt{\frac{(6-1)\times 5.53 + (10-1)\times 3.12}{(6-1) + (10-1)}} = 1.995.$$

M-2.1.3.5. Calculate

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{s_E \sqrt{1/n + 1/m}} = \frac{5.87 - 4.61 - 0}{1.995\sqrt{1/10 + 1/6}} = 1.22 .$$

M-2.1.3.6. Because we want an 80% level of confidence, $\alpha = 0.20$. So, $t_{0.80,14} = 0.8681$. Now compare the calculated value, *t*, with the critical value, $t_{0.80,14}$: $1.22 > 0.8681$. Therefore, reject H_0. At the 80% level of confidence, the mean concentration of Ni at Site A is greater than the mean background concentration of Ni.

M-2.2. Satterthwaite's t-Test (Unequal Variances). If the two variances are not equal, the use of Satterthwaite's *t*-test is recommended. Directions are provided below in Paragraph M-2.2.1, followed by an example in Paragraph M-2.2.2.

M-2.2.1. Directions for Applying Satterthwaite's t-Test to Unequal Variances. This describes the steps for applying the two-sample *t*-test for differences between the population means for: Case 1: $H_0 : \mu_X - \mu_Y \le \delta_o$ vs. $H_A : \mu_x - \mu_y > \delta_0$; and Case 2: $H_0 : \mu_x - \mu_y \ge \delta_0$ vs. $H_A : \mu_x - \mu_y < \delta_0$, which is given in braces { }.

M-2.2.1.1. Verify that both data sets come from a normal distribution, using the tests presented in Appendices F and J, such as the Shapiro-Wilk test (Paragraph F-3.2) and a normal probability plot (Paragraph J-5.5).

M-2.2.1.2. Calculate the sample mean, $\bar{x}$, and the sample variance, s_X^2 (Appendix C), for sample 1 and compute the sample mean, $\bar{y}$, and the sample variance, s_Y^2, for sample 2.

M-2.2.1.3. Test for equal variances, using tests presented in Appendix N, such as Bartlett's test (Paragraph N-3). If the variances are approximately equal, use the two-sample *t*-test (presented in Paragraph M-2.2.2). Otherwise, compute the standard deviation for unequal variances:

$$s_{NE} = \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}} .$$

M-2.2.1.4. Calculate

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{s_{NE}} .$$

M-2.2.1.5. Use Table B-23 of Appendix B to find the critical value, $t_{1-\alpha,\nu}$, such that $100(1-\alpha)$% of the *t*-distribution with *v* degrees of freedom is below $t_{1-\alpha,\nu}$, and

$$\nu = \frac{\left[\dfrac{s_X^2}{m} + \dfrac{s_Y^2}{n}\right]^2}{\dfrac{s_x^2}{m^2(m-1)} + \dfrac{s_y^2}{n^2(n-1)}} \ .$$

Round down the degrees of freedom to the nearest integer. Compare *t* to the critical value:

M-2.2.1.5.1. If $t > t_{1-\alpha,\nu}\{t < -t_{1-\alpha,\nu}\}$, H_0 may be rejected.

M-2.2.1.5.2. If $t \le t_{1-\alpha,\nu}\{t \ge -t_{1-\alpha,\nu}\}$, there is not enough evidence to reject H_0. Therefore, the false acceptance error rate will need to be verified. Go to M-2.2.1.6.

M-2.2.1.6. If H_0 was not rejected, calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance 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 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.*

M-2.2.1.7. The results of the test could be:

M-2.2.1.7.1. H_0 is rejected: $\mu_x - \mu_y > \delta_0\{\mu_x - \mu_y < \delta_0\}$.

M-2.2.1.7.2. H_0 is not rejected and the false acceptance error rate is satisfied, $\mu_x - \mu_y \le \delta_0\{\mu_x - \mu_y \ge \delta_0\}$.

M-2.2.1.7.3. H_0 is not rejected but the false acceptance error rate is not satisfied; H_0 is uncertain because the sample size was too small.

M-2.2.2. Example of Applying Satterthwaite's t-test to Unequal Variances. Because we want a 95% level of confidence, $\alpha = 0.05$ and *v* = 6 (round down to the nearest integer). So, $t_{0.95,6} = 1.943$. Now compare the calculated value (*t*) with the critical value, $t_{0.95,6}$. Because $-1.031 \le 1.943$, there is not enough evidence to reject H_0.

M-2.2.2.1. As a result of not having enough evidence to reject the null hypothesis, it is necessary to calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates. DEFT can be used to evaluate power and sample size and is presented in this example. To calculate the power of the test, one must consider what an acceptable difference among the means is before concluding H_0 should be rejected. The difference that one is willing to accept depends on the detection limits achieved, the range of concentrations from each data set, and what is considered to have practical significance vs. statistical significance.

M-2.2.2.2. The power curve (Figure M-1) shows where a statistically significant difference between the means was assumed to be 1 mg/kg (the region between the vertical dashed and solid lines). According to DEFT, 21 samples are needed for the estimated performance curve. In the above example, the site data have 36 samples and the background data only have 8. Therefore, there may be a need to take more background samples. It is important to note that the true difference in the mean (4.619 – 4.925 = –0.31) is to the left of the action level.

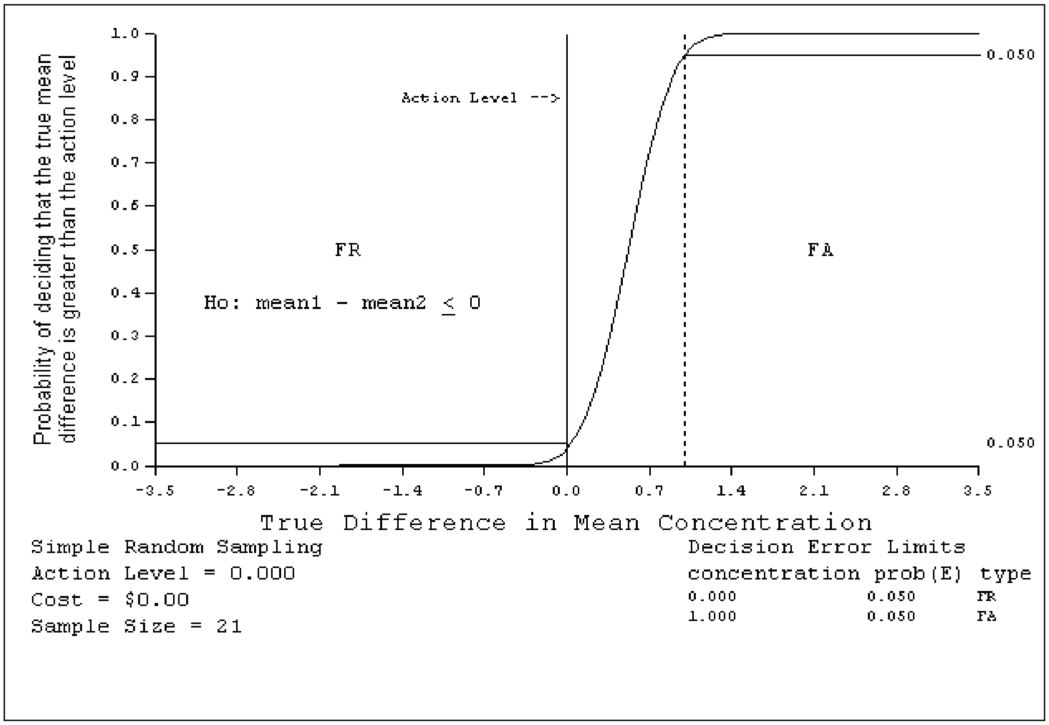


Figure M-1. Estimated Power Performance Curve.

M-2.3. Matched Pairs t-Test.

M-2.3.1. Introduction. Sometimes, the two populations of interest represent different measurements on the same homogenous group. For example, contaminant concentration in groundwater before and after a certain remediation treatment may need to be compared. If measurements are taken from the same set of wells both before and after treatment, we can match the results by well. That is, each well will have a result from before the treatment and a result from after the treatment. Under this experimental design, the observed differences for each well before and after treatment become the sample data because we expect the two results from each well to be more homogeneous than the results among wells.

M-2.3.1.1. The differences are then analyzed using the one-sample *t*-test if the assumptions for that test are met. Namely, the one-sample *t*-test assumes the differences represent a random sample. It also assumes that the average difference follows a normal distribution. If the normal assumption is not valid, Paragraph M-4.1.6 discusses a non-parametric alternative for matched pairs designs. In addition to matched pairs, one would ideally assign the order of the treatments randomly to each subject, although that would not be possible in the groundwater remediation example. Matching can also occur between subjects that are closely alike in all respects except the treatment that is applied.

M-2.3.1.2. The matched pairs *t*-test is commonly used to compare site contaminant concentrations before and after a treatment:

$$H_0 : \mu_A \geq \mu_B,\ H_A : \mu_A < \mu_B\ .$$

M-2.3.1.3. The “true” mean concentration *before* treatment and the “true” mean concentration *after* treatment are denoted by μ_B and μ_A, respectively. The before treatment mean is often referred to as the “baseline” mean. Directions are provided below in Paragraph M-2.3.2, followed by an example in Paragraph M-2.3.3.

M-2.3.2. Directions to Apply the Matched Pairs t-test for Differences Between the Means Before and After a Treatment. Steps to apply the Matched Pairs *t*-test for differences between the means for Case 1 and Case 2 are as follows: Case 1: $H_0 : \mu_A \geq \mu_B$, $H_A : \mu_A < \mu_B$; and Case 2: $H_0 : \mu_A \leq \mu_B$, $H_A : \mu_A > \mu_B$, which is given in braces { }.

M-2.3.2.1. Subtract the before treatment concentration (B_i) from the corresponding after treatment concentration (A_i) for each pair of results (B_i, A_i) to obtain the differences:

$$d_i = A_i - B_i\ .$$

M-2.3.2.2. Verify that the differences, d_1, d_2, $d_3 ... d_n$, are normal, using procedures in Appendices F and J, such as the Shapiro-Wilk test (Paragraphs F-3.2 and F-3.3) and a normal probability plot (Paragraphs J-5.5 and J-5.6).

M-2.3.2.3. Calculate the sample mean, $\bar{d}$, and the sample variance, s_d^2 (Appendix D).

M-2.3.2.4. Calculate

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}\ .$$

M-2.3.2.5. Use Table B-23 of Appendix B to find the critical value, $t_{1-\alpha,n-1}$, such that $(1-\alpha)100\,\%$ of the *t* distribution with (n – 1) degrees of freedom is below $t_{1-\alpha,n-1}$.

M-2.3.2.5.1. If $t < -t_{1-\alpha,n-1}\{t > t_{1-\alpha,n-1}\}$, reject H_0. Go to M-2.3.2.7.

M-2.3.2.5.2. If $t \geq -t_{1-\alpha,n-1}\{t \leq t_{1-\alpha,n-1}\}$, there is not enough evidence to reject H_0. Therefore, the false acceptance error rate will need to be verified. Go to M-2.3.2.6.

M-2.3.2.6. 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 matched pairs *t*-test. If only one false acceptance error rate (β) has been specified (at μ_1), it is possible to approximately 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. A derivation of the following formula is given in Appendix A of EPA 600/R-96/055, QA/G-4.

M-2.3.2.7. Calculate:

$$m = \frac{s_d^2\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{\left(\bar{d}\right)^2} + (0.5)Z_{1-\alpha}^2$$

where Z_p is the $p100^{\text{th}}$ percentile of the standard normal distribution (Table B-15 of Appendix B). Round *m* up to the next integer. If $m \leq n$, the false acceptance error rate has been satisfied. If $m > n$, the false acceptance error rate has not been satisfied.

M-2.3.2.8. The results of the test could be:

M-2.3.2.8.1. H_0 is rejected; $\mu_A < \mu_B\,\{\mu_A > \mu_B\}$.

M-2.3.2.8.2. H_0 is not rejected and the false acceptance error rate is satisfied; $\mu_A \geq \mu_B\,\{\mu_A \leq \mu_B\}$.

M-2.3.2.8.3. H_0 is not rejected and the false acceptance error rate was not satisfied; $\mu_A \geq \mu_B\,\{\mu_A \leq \mu_B\}$, but this conclusion is uncertain because the sample size was too small.

M-2.3.3. Example of the Matched Pairs t-Test for the Difference Between Means Before and After Treatment. Consider the case where the results of a groundwater remediation procedure are compared before and after treatment to determine if the remediation has decreased the concentration of the contaminant. Test the null hypothesis that the treatment had no lowering effect at the 95% level of confidence:

$$H_0 : \mu_A \geq \mu_B,\ \ H_A : \mu_A < \mu_B\ .$$

M-2.3.3.1. The data consist of measured TCE concentrations (mg/L) at monitoring wells before and after a treatment-test, given in Table M-1.

M-2.3.3.2. Determine if the differences follow a normal distribution. A Shapiro-Wilk test for normality does not reject the hypothesis that the differences are normal (*p* = 0.4248). So, assuming normality is reasonable.

M-2.3.3.3. Calculate

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{-18.0}{13.9/\sqrt{10}} = -4.10\,.$$

M-2.3.3.4. Assume that we want a 95% level of confidence, $\alpha = 0.05$. So, $t_{0.95,9} = 1.833$. Now compare the calculated value, *t*, with the critical value $-t_{0.95,9}$: $-4.10 < -1.833$. Therefore, reject H_0. This means that there is a lower mean concentration of TCE after remediation.

Table M-1.
TCE Concentrations (mg/L) at Monitoring Wells Before and After a Treatment

Sample ID	Baseline (01/2000)	Post–Test (12/2000)	Difference
Well 1	20.9	0.917	–20.0
Well 2	9.17	8.77	–0.400
Well 3	5.96	4.37	–1.59
Well 4	41.5	4.34	–37.2
Well 5	34.3	10.7	–23.6
Well 6	19.7	1.48	–18.2
Well 7	38.9	0.272	–38.6
Well 8	8.18	0.520	–7.66
Well 9	9.13	3.06	–6.07
Well 10	28.5	1.90	–26.6

M-3. <u>Comparing Proportions and Percentiles: Two-Sample Test for Proportions</u>. This Paragraph considers hypotheses concerning two population proportions (or percentiles). 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 of size *n* ($y_1, y_2, \ldots, y_n$) from the second population. The sample proportion for the first population is represented by p_1 and the sample proportion for the second population is represented by p_2.

M-3.1. <u>Introduction</u>. The principal assumption for this non-parametric test is that of random sampling from the two populations. 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. Directions for a two-sample test for proportions for a simple random sample and a systematic simple random sample are given below in Paragraph M-3.2, followed by an example in Paragraph M-3.3.

M-3.2. Directions for Applying the Two-Sample Test for Proportions. Directions for applying the two-sample test for proportions are presented for Case 1: $H_0 : P_1 - P_2 \leq \delta_0$ and $H_A : P_1 - P_2 > \delta_0$; and Case 2: $H_0 : P_1 - P_2 \geq \delta_0$ and $H_A : P_1 - P_2 < \delta_0$, which is given in braces { }. 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 some concentration *C*, and let k_2 be the number of points from sample 2 that exceed *C*.

M-3.2.1. Calculate the sample proportions: $p_1 = k_1 / m$, $p_2 = k_2 / n$.

M-3.2.2. Calculate the pooled proportion: $p = (k_1 + k_2)/(m + n)$.

M-3.2.3. Compute:

mp_1, $m(1 - p_1)$, np_2, $n(1 - p_2)$.

If all of the above values are greater than or equal to 5, continue. Otherwise, seek assistance from a statistician as analysis is complicated.

M-3.2.4. Calculate:

$$z = (p_1 - p_2) / \sqrt{p(1-p)(1/m + 1/n)}$$

M-3.2.5. Use Table B-15 of Appendix B to find the critical value, $Z_{1-\alpha}$, such that $(1-\alpha)100\%$ of the normal distribution is below $Z_{1-\alpha}$. For example, if α = 0.05 then $Z_{1-\alpha}$ = 1.645.

M-3.2.5.1. If $z > Z_{1-\alpha}$ $\{z < -Z_{1-\alpha}\}$, reject H_0.

M-3.2.5.2. If $z \leq Z_{1-\alpha}$ $\{z \geq -Z_{1-\alpha}\}$, do not reject H_0. Proceed to M-3.2.6 to calculate the false acceptance error rate.

M-3.2.6. If H_0 is not rejected, calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates. If only one false acceptance 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:

$$m^* = n^* = \frac{2\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2 \overline{P}\,(1-\overline{P})}{\left(P_2 - P_1\right)^2}$$

$$\overline{P} = \frac{P_1 + P_2}{2} \ .$$

Z_p is the $p100^{\text{th}}$ percentile of the standard normal distribution (Table B-15 of Appendix B).

M-3.2.6.1. If $m > m^*$ and $n > m^*$, then the false acceptance error rate has been satisfied.

M-3.2.6.2. If both m and n are below m^*, the false acceptance error rate has not been satisfied.

M-3.2.6.3. If m^* is between m and n, use a software package like the DEFT or DataQUEST 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.

M-3.2.6.4. If the estimated power is below $1 - \beta$, the false acceptance error rate has not been satisfied.

M-3.2.7. The results of the test could be:

M-3.2.7.1. H_0 is rejected; $P_1 - P_2 > \delta_0$ { $P_1 - P_2 < \delta_0$ }.

M-3.2.7.2. H_0 was not rejected, the false acceptance error rate was satisfied, and it seems $P_1 - P_2 \le \delta_0$ { $P_1 - P_2 \ge \delta_0$ }.

M-3.2.7.3. H_0 was not rejected, the false acceptance error rate was not satisfied, and it seems $P_1 - P_2 \le \delta_0$ { $P_1 - P_2 \ge \delta_0$ }, but this outcome is uncertain because the sample size was probably too small.

M-3.3. Example of Two-Sample Test for Proportions for Simple and Systematic Random Samples. Gasoline groundwater concentrations at Site A are compared to background concentrations:

$$H_0 : P_1 - P_2 \le \delta_0 \,,\ \ H_A : P_1 - P_2 > \delta_0 \ .$$

M-3.3.1. The groundwater site data are following (m = 15): 243, 700, 781, 385, 642, 97.2, 233, 11.1, 10.60, 14.90, 14.90, 12.70, 9.57, 6.04, and 7.32 µg/L.

M-3.3.2. The groundwater background data are following (n = 45): 177.0, 4.27, 10.60, 10.60, 14.90, 14.60, 12.70, 9.57, 95.70, 7.32, 7.32, 7.32, 6.58, 6.90, 6.90, 39.5, 4.27, 10.60, 10.60, 14.90, 14.60, 12.70, 9.57, 6.04, 7.32, 7.32, 7.32, 146.00, 6.90, 6.90, 44.5, 4.27, 10.60, 10.60, 14.90, 14.60, 12.70, 9.57, 6.04, 7.32, 7.32, 7.32, 111.00, 6.90, and 6.90 µg/L.

Data	k_i	**Sample Size**
Site (*i* = 1)	7	15
Background (*i* = 2)	6	45

where k_i is the number of detected concentrations above the regulatory threshold (35 µg/L).

M-3.3.3. Determine whether or not mp_1, $m(1-p_1)$, np_2, $n(1-p_2)$ are all greater than 5:

$$p_1 = k_1/m = 7/15 = 0.467$$

$$p_2 = k_2/n = 6/45 = 0.133$$

$$mp_1 = 15(0.467) = 7 > 5$$

$$m(1-p_1) = 15(1-0.467) = 8 > 5$$

$$np_2 = 45(0.133) = 6 > 5$$

$$n(1-p_2) = 45(1-0.133) = 39 > 5.$$

M-3.3.4. Calculate the following:

$$p = (k_1 + k_2)/(m+n) = (7+6)/(15+45) = 0.217$$

$$\begin{aligned} z &= (p_1 - p_2)/\sqrt{p(1-p)(1/m + 1/n)} \\ &= (0.467 - 0.133)/\sqrt{0.217(1-0.217)(1/15 + 1/45)} = 2.72 \end{aligned} .$$

M-3.3.5. Because the level of confidence is 95%, $\alpha = 0.05$. Using Table B-15, we find that $Z_{1-0.05} = 1.645$. Now compare the calculated value, z, with the critical value, $Z_{1-0.05}$: $2.74 > 1.645$.

M-3.3.6. Therefore, there is enough evidence to reject H_0 (i.e., the results suggest that the proportion of samples with gasoline levels above the regulatory threshold in the site well samples is greater than the proportion above the regulatory threshold in the background well samples).

M-4. Nonparametric Comparisons of Two Populations.

M-4.1. The Wilcoxon Rank Sum Test. The Wilcoxon rank sum test is a nonparametric test that can be used to compare two population distributions based on *n* independent random samples ($x_1, x_2, \ldots, x_n$) from the first population, and *m* independent random samples ($y_1, y_2, \ldots, y_m$) from the second population. The most general form of the hypotheses for a one-tailed Wilcoxon rank sum test can be stated in terms of the probability that an observation from distribution *Y* exceeds a value from distribution *X*, such as:

$$H_0 : P(X < Y) \geq 0.5, \quad H_A : P(X < Y) < 0.5 \ .$$

M-4.1.2. Introduction. Hypotheses on the relative rank of the mean of each population can also be formulated with the additional assumption that the two underlying distributions have the same shape and dispersion (Conover, 1980). That is, one distribution differs by some fixed amount (or is increased by a constant) when compared to the other distribution. An important advantage of the Wilcoxon rank sum test is its partial robustness to outliers, because the analysis is conducted on rankings of the observations. This limits the influence of outliers because a given observation can be no more extreme than the first or last rank. Directions and an example for the Wilcoxon rank sum test are given in Paragraphs M-4.1.3 and M-4.1.4, respectively. If a relatively large number of samples have been taken, it is more efficient to use the large sample approximation to the Wilcoxon rank sum test (Paragraph M-4.1.6) to perform the hypothesis test.

M-4.1.3. Directions for the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples.

M-4.1.3.1. Let $x_1, x_2, \ldots, x_n$ represent the *n* observations from population 1 and $y_1, y_2, \ldots, y_m$ represent the *m* observations from population 2, where both *n* and *m* are less than or equal to 20.

M-4.1.3.1.1. *Case 1*:

$H_0 : P(X < Y) \geq 0.5$: Values of *X* tend to be smaller than or equal to values of *Y*.

$H_A : P(X < Y) < 0.5$: Values of *X* tend to be larger than values of *Y*.

M-4.1.3.1.2. *Case 2*:

$H_0 : P(X < Y) \leq 0.5$: Values of *X* tend to be larger than or equal to values of *Y*.

$H_A : P(X < Y) > 0.5$: Values of *X* tend to be smaller than values of *Y*.

M-4.1.3.1.3. *Case 3*:

$H_0 : P(X < Y) = 0.5$: Values of *X* tend to be equal to values of *Y*.

$H_A : P(X < Y) \neq 0.5$: Values of *X* tend to be smaller than or greater than values of *Y*.

M-4.1.3.2. If either *m* or *n* is larger than 20 and the smaller of the two is at least 4 (Lehmann, 1975), use the large sample approximation described in Paragraph M-4.1.5.

M-4.1.3.3. Combine the two data sets and rank the measurements (from both data sets) from smallest to largest, keeping track of which population contributed each measurement.

M-4.1.3.3.1. Assign the rank of 1 to the smallest value of the combined data sets and note whether the smallest value is from population 1 or 2.

M-4.1.3.3.2. Assign the rank of 2 to the second smallest value of the combined data sets (noting the population), and so forth.

M-4.1.3.3.3. If there are ties, assign the average of the ranks that would otherwise have been assigned to the tied observations.

M-4.1.3.4. Calculate *R*, the sum of the ranks of the data from population 1, and then calculate:

$$W = R - \frac{n(n+1)}{2}.$$

M-4.1.3.5. Use Table B-17 of Appendix B to find the critical value, W_α (or $W_{\alpha/2}$ for Case 3).

M-4.1.3.6. Compare *W* to the critical value W_α.

M-4.1.3.6.1. For Case 1, reject H_0 if $W > nm - W_\alpha$.

M-4.1.3.6.2. For Case 2, reject H_0 if $W < W_\alpha$.

M-4.1.3.6.3. For Case 3, reject H_0 if $W > nm - W_{\alpha/2}$ or $W < W_{\alpha/2}$.

M-4.1.3.7. The results of the test could be:

M-4.1.3.7.1. H_0 was rejected and it seems values from population 1 tend to be greater than (Case 1), smaller than (Case 2), or different from (Case 3) values from population 2.

M-4.1.3.7.2. H_0 was not rejected, and it seems that values from population 1 tend to be smaller than or equal to (Case 1), greater than or equal to (Case 2), or not different from (Case 3) values from population 2.

M-4.1.3.7.3. If H_0 is not rejected, it should be determined whether adequate power was achieved. However, as power calculations tend to be complex and difficult to do manually, it is recommended that a statistician be consulted.

M-4.1.4. Example of the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples.

M-4.1.4.1. Consider the Case 1 (Paragraph M-4.1.3), where lead (Pb) surface soil concentrations are compared between Site A and background at a significance level of $\alpha =$ 0.05 using the test.

M-4.1.4.1.1. H_0: Site A Pb concentrations tend to be less than or equal to background Pb concentrations.

M-4.1.4.1.2. H_A: Site A Pb concentrations tend to be greater than background Pb concentrations.

M-4.1.4.2. Suppose the Pb surface site concentrations (*X*) are as follows (*n* = 20): 8.24, 6.57, 4.48, 4.34, 16.00, 3.83, 4.11, 3.48, 3.66, 5.01, 93.80, 3.70, 129.00, 4.92, 91.80, 3.86, 4.21, 4.32, 10.00, and 9.38 mg/kg.

M-4.1.4.3. Suppose the Pb surface background concentrations (*Y*) are as follows (*m* = 16): 3.81, 3.68, 3.72, 3.68, 5.97, 4.12, 6.42, 4.13, 8.88, 3.01, 5.34, 3.74, 10.70, 3.86, 10.80, and 4.40 mg/kg.

$$W = R - \frac{n(n+1)}{2} = 409.5 - \frac{20(20+1)}{2} = 199.5$$

$$W_\alpha = W_{0.05} = 108$$

$$nm - W_\alpha = (20)(16) - 108 = 212 \ .$$

M-4.1.4.4. Because $199.5 \le 212$, H_0 cannot be rejected. There is insufficient evidence to conclude that the lead concentrations from Site A are greater than background lead concentrations.

Table M-2.
Example M-4.1.4 Pb Concentrations

Location	Result (mg/Kg)	Rank	Location	Result (mg/Kg)	Rank
background	3.01	1	background	4.4	19
Site	3.48	2	site	4.48	20
Site	3.66	3	site	4.92	21
background	3.68	4.5	site	5.01	22
background	3.68	4.5	background	5.34	23
Site	3.70	6	background	5.97	24
background	3.72	7	background	6.42	25
background	3.74	8	site	6.57	26
background	3.81	9	site	8.24	27
Site	3.83	10	background	8.88	28
background	3.86	11.5	site	9.38	29
Site	3.86	11.5	site	10.0	30
Site	4.11	13	background	10.7	31
background	4.12	14	background	10.8	32
background	4.13	15	site	16.0	33
Site	4.21	16	site	91.8	34
Site	4.32	17	site	93.8	35
Site	4.34	18	site	129.0	36

M-4.1.5. Large Sample Approximation of the Wilcoxon Rank Sum Test. When a relatively large number of samples has been taken, it is more efficient to use a large sample approximation of the Wilcoxon rank sum test to obtain the critical value of *W*. Directions and an example are presented in Paragraphs M-4.1.5.1 and M-4.1.5.2, respectively. Required sample size to achieve a specified power is explored in Paragraphs M-4.1.4.3 and M-4.1.4.4.

M-4.1.5.1. Directions for a Large Sample Approximation of the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples.

M-4.1.5.1.1. Let $x_1, x_2, \ldots, x_n$ represent the *n* observations from population 1 and $y_1, y_2, \ldots, y_m$ represent the *m* observations from population 2 where either *n* or *m* is greater than 20 and the smaller of *n* and *m* is at least 4 (Lehmann, 1975). The following hypothesis tests are considered:

M-4.1.5.1.1.1. *Case 1.* $H_0 : P(X < Y) \geq 0.5$, $H_A : P(X < Y) < 0.5$.

M-4.1.5.1.1.2. *Case 2.* $H_0 : P(X < Y) \leq 0.5$, $H_A : P(X < Y) > 0.5$.

M-4.1.5.1.1.3. *Case 3.* $H_0 : P(X < Y) = 0.5$, $H_A : P(X < Y) \neq 0.5$.

M-4.1.5.1.2. List and rank the measurements from both populations from smallest to largest, keeping track of which population contributed each measurement.

M-4.1.5.1.2.1. The rank of 1 is assigned to the smallest value of the combined data sets, the rank of 2 to the second smallest value of the combined data sets, and so forth.

M-4.1.5.1.2.2. If there are ties, assign the average of the ranks that would otherwise have been assigned to the tied observations.

M-4.1.5.1.3. Calculate *R,* the sum of the ranks of the data from population 1, and then calculate:

$$W = R - \frac{n(n+1)}{2} .$$

M-4.1.5.1.4. Calculate:

$$w_p = \frac{mn}{2} + Z_p \sqrt{mn(n+m+1)/12} .$$

M-4.1.5.1.4.1. *Case 1*. $p = 1 - \alpha$

M-4.1.5.1.4.2. *Case 2*: $p = \alpha$

M-4.1.5.1.4.3. *Case 3*. Calculate both $w_{\alpha/2}(p = \alpha/2)$ and $w_{1-\alpha/2}(p = 1 - \alpha/2)$ (Lehmann, 1975).

M-4.1.5.1.5. Note that Z_p is the $p100^{th}$ percentile of the standard normal distribution (Table B-15 of Appendix B).

M-4.1.5.1.5.1. For Case 1, reject H_0 if $W > w_{1-\alpha}$.

M-4.1.5.1.5.2. For Case 2, reject H_0 if $W < w_{\alpha}$.

M-4.1.5.1.5.3. For Case 3, reject H_0 if $W > w_{1-\alpha/2}$ or $W < w_{\alpha/2}$.

M-4.1.5.1.6. The results of the test could be as follows.

M-4.1.5.1.6.1. H_0 was rejected and it seems values from population 1 tend to be greater than (Case 1), smaller than (Case 2), or different from (Case 3) values from population 2.

M-4.1.5.1.6.2. H_0 was not rejected, and it seems that values from population 1 tend to be smaller than or equal to (Case 1), greater than or equal to (Case 2), or not different from (Case 3) values from population 2.

M-4.1.5.2. Example of the Large Sample Approximation to the Wilcoxon Rank Sum Test for Simple and Systematic Random Samples.

M-4.1.5.2.1. Consider the case where lead (Pb) surface soil concentrations are compared between Site A and background at a significance level of 0.05 using the test (Case 1 in Paragraph M-4.1.5.1) (Table M-3).

M-4.1.5.2.1.1. H_0: Site A Pb concentrations tend to be less than or equal to background Pb concentrations.

M-4.1.5.2.1.2. H_A: Site A Pb concentrations tend to be larger than background lead concentrations.

Table M-3.
Example M-4.1.5.2 Pb Concentrations

Location	Result	Rank	Location	Result	Rank
Background	3.01	1	site	4.48	22
Background	3.05	2	site	4.92	23
Site	3.48	3	site	5.01	24
Site	3.66	4	background	5.34	25
Background	3.68	5.5	background	5.97	26
Background	3.68	5.5	background	6.2	27
Site	3.7	7	background	6.42	28
Background	3.72	8	site	6.57	29
Background	3.74	9	site	8.24	30
Background	3.81	10	background	8.88	31
Site	3.83	11	site	9.38	32
Site	3.86	12.5	site	10	33
Background	3.86	12.5	background	10.7	34
Site	4.11	14	background	10.8	35
Background	4.12	15	background	15.5	36
Background	4.13	16	site	16	37
Background	4.2	17	background	20.6	38
Site	4.21	18	site	91.8	39
Site	4.32	19	site	93.8	40
Site	4.34	20	site	129	41
Background	4.4	21	—	—	—

M-4.1.5.2.2. Suppose the surface soil Pb concentrations for Site A (X) are: 8.24, 6.57, 4.48, 4.34, 16.00, 3.83, 4.11, 3.48, 3.66, 5.01, 93.80, 3.70, 129.00, 4.92, 91.80, 3.86, 4.21, 4.32, 10.00, and 9.38 mg/kg.

M-4.1.5.2.3. Suppose the background surface soil Pb concentrations (*Y*) are: 3.05, 3.81, 3.68, 3.72, 4.20, 3.68, 5.97, 4.12, 6.42, 6.20, 4.13, 8.88, 3.01, 15.5, 5.34, 3.74, 20.6, 10.70, 3.86, 10.80, and 4.40 mg/kg.

M-4.1.5.2.4. Note that tied values occur at for concentrations 3.68 and 3.86. These ties are assigned the average of the ranks they would otherwise have been assigned. The rank of 3.68 is 5.5, which is the average of ranks 5 and 6, and the rank of 3.86 is 12.5, which is the average of ranks 12 and 13.

M-4.1.5.2.5. Population 1 is the lead surface site data (n = 20), and population 2 is the background lead data (m = 21). Calculate W as:

$$W = R - \frac{n(n+1)}{2} = 458.5 - \frac{20(20+1)}{2} = 248.5.$$

M-4.1.5.2.6. Calculate

$$w_p = \frac{mn}{2} + Z_p\sqrt{mn(n+m+1)/12} = \frac{21\times 20}{2} + 1.645\sqrt{21\times 20(20+21+1)/12} = 273.1$$

$$Z_p = Z_{1-\alpha} = Z_{0.95} = 1.645.$$

M-4.1.5.2.6. Compare the calculated statistic W to the critical value $w_{1-\alpha}$, (248.5 < 273.1). Because $W \le w_{1-\alpha}$, do not reject the null hypothesis. Lead concentrations from Site A may be less than or equal to background lead concentrations. The power of the test needs to be determined (refer to Paragraph M-4.1.5.3).

M-4.1.5.3. Directions for Calculating Sample Size to Achieve a Specified Power for the Wilcoxon Rank Sum Test.

M-4.1.5.3.1. Noether (1987) discusses the determination of an adequate sample size based on a predefined level of power to apply the Wilcoxon rank sum test for the following hypothesis test. The n values of X ($x_1, x_2, \ldots, x_n$) compared to m values of Y ($y_1, y_2, \ldots, y_m$):

M-4.1.5.3.1.1. Case 1. $H_0 : P(X < Y) \ge 0.5$, $H_A : P(X < Y) < 0.5$.

M-4.1.5.3.1.2. Case 2. $H_0 : P(X < Y) \le 0.5$, $H_A : P(X < Y) > 0.5$.

M-4.1.5.3.1.3. Case 3: $H_0 : P(X < Y) = 0.5$, $H_A : P(X < Y) \ne 0.5$.

M-4.1.5.3.2. The total number of samples collected, $N = n + m$, is compared with a conservative estimate of the number of samples N' required to achieve some desired power $1 - \beta$ Under the assumption that the test statistic (in this case, the large sample approximation for the Wilcoxon rank sum statistic in Paragraph M-4.1.5.1) is normally distributed, N' is determined as follows. For Cases 1 and 2:

$$N' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{12c\left(1-c\right)\left(p'' - \frac{1}{2}\right)^2}$$

and for Case 3:

$$N' = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{12c\left(1-c\right)\left(p'' - \frac{1}{2}\right)^2}$$

where Z_q = q quantile of the standard normal distribution (from Table B-15)

α = significance level of the test

$1-\beta$ = desired power for the test

$$c = \frac{n}{N}$$

$p'' =$ $P\left(X < Y\right)$.

M-4.1.5.3.4. Setting c equal to 0.5 will be best unless there are reasons to sample more heavily from one of the populations. The value of p'' can be taken from past information, a pilot sample, or chosen to represent a meaningful shift in the data (Noether, 1987). The normality of the test statistic under the null hypothesis is generally valid if either n or m exceeds 20 and the smaller of the two is at least 4. If the suggested sample size does not meet these requirements, consult a statistician.

M-4.1.5.4. Example of Calculating Sample Size to Achieve a Specified Power for the Wilcoxon Rank Sum Test. Suppose Pb surface soil concentrations at a site are to be compared to background concentrations using a 95% level of confidence ($\alpha = 0.05$) using the following hypothesis test (Case 1).

M-4.1.5.4.1. H_0: Site A Pb concentrations tend to be less than or equal to background concentrations.

M-4.1.5.4.2. H_A: Site A Pb concentrations tend to be higher than background concentrations.

M-4.1.5.4.3. We wish to ensure that the sample size is large enough to find a meaningful elevation of lead concentrations with 80% probability ($\beta = 0.20$). Suppose historical information indicates that the probability of site lead concentration being less than background lead concentration is about 1/3. We decide to use this as our estimate of p''. We wish to take an equal number of samples from the site and background, so that *c* = 0.5. The required sample size to meet the power requirement is:

$$N' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{12c\left(1-c\right)\left(p'' - \frac{1}{2}\right)^2} = \frac{\left(1.645 + 0.842\right)^2}{12\left(0.5\right)\left(1-0.5\right)\left(0.333-0.5\right)^2} = 74.2\ .$$

M-4.1.5.4.4. As we wish to collect and equal number of samples from the site and background, the calculated required total sample size is rounded up to the next largest even whole number, 76 (an even number is required because it is being assumed that the required sample size is equal to the sum of an equal number of site and background samples). If it is assumed that 38 site plus 38 background samples are required to achieve adequate power for the test performed in Paragraph M-4.1.5.2, it follows that, though the null hypothesis was not rejected, the result is not conclusive (as only 20 site and 21 background samples were collected).

M-4.1.6. Matched Pairs Wilcoxon Signed Ranks Test. As discussed in Paragraph M-2.3, matching subjects can lead to efficient comparisons between two populations. However, the observed differences between treatments will not always appear to come from a normal distribution. In that case, the one-sample Wilcoxon signed ranks test that was discussed in Appendix L can be used to test whether the mean or median difference differs significantly from zero. Directions for applying the Wilcoxon signed ranks test to a matched pairs design are presented in Paragraph M-4.1.6.1 and an example is presented in Paragraph M-4.1.6.2. See the discussion in Appendix L for more details on applying the Wilcoxon signed ranks test.

M-4.1.6.1. Directions for the Wilcoxon Signed Ranks Test for Matched Pairs. The following describes the steps for applying the Wilcoxon signed ranks test for a matched pairs design when the sample size, *n*, is less than 20 for: Case 1: $H_0 : \mu_A \geq \mu_B$, $H_A : \mu_A < \mu_B$; and Case 2: $H_0 : \mu_A \leq \mu_B$, $H_A : \mu_A > \mu_B$, which is given in braces { }.

M-4.1.6.1.1. Subtract each before concentration (B_{i}) from the after concentration (A_{i}) to get the difference:

$$d_i = A_i - B_i\ .$$

If any of the differences are zero, delete them and correspondingly reduce the sample size (*n*).

M-4.1.6.1.2. Assign ranks from 1 to *n* based on ordering the absolute deviations $|d_i|$ (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 that would otherwise have been assigned to the tied observations.

M-4.1.6.1.3. Assign the sign for each observation to create the signed rank. The sign is positive if the deviation d_i is positive and the sign is negative if the deviation d_i is negative.
Calculate *R*, the sum of the ranks with a positive sign.

M-4.1.6.1.4. Use Table B-24 of Appendix B to find the critical value $w_{\alpha,n}$.

M-4.1.6.1.5. Compare the calculated test statistic, *R*, to the critical value:

M-4.1.6.1.5.1. If $R \le \{n(n+1)/2\} - w_{\alpha,n}\ \{R \ge w_{\alpha,n}\}$, H_0 may be rejected.

M-4.1.6.1.5.2. If $R > \{n(n+1)/2\} - w_{\alpha,n}\ \{R < w_{\alpha,n}\}$, there is not enough evidence to reject H_0; verify the false acceptance error rate.

M-4.1.6.1.6. If H_0 was not rejected, calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates using a software package like DEFT (EPA QA/G-4D).

M-4.1.6.1.7. The results of the test may be:

M-4.1.6.1.7.1. H_0 is rejected; $\mu_A < \mu_B\ \{\mu_A > \mu_B\}$.

M-4.1.6.1.7.2. H_0 is not rejected and the false acceptance error rate is satisfied; $\mu_A \ge \mu_B\ \{\mu_A \le \mu_B\}$.

M-4.1.6.1.7.3. H_0 is not rejected and the false acceptance error rate was not satisfied; $\mu_A \ge \mu_B\ \{\mu_A \le \mu_B\}$, but this conclusion is uncertain because the sample size was too small.

M-4.1.6.2. Example of the Matched Pairs Wilcoxon Signed Ranks Test for the Difference Between Means Before and After Treatment. Consider the case where the results of a groundwater remediation procedure are compared before and after treatment to see if the remediation has lowered the concentration of the contaminant. Test the hypothesis that the treatment had no lowering effect at the 95% level of confidence:

$H_0 : \mu_A \geq \mu_B$, $H_A : \mu_A < \mu_B$.

M-4.1.6.2.1. The data consist of measured TCE concentrations (mg/L) at monitoring wells before and after treatment (Table M-4). Negative values of the difference support the alternative hypothesis.

M-4.1.6.2.2. The differences are roughly symmetrical so the Wilcoxon signed ranks test can be applied.

M-4.1.6.2.3. Because the sign ranks are all negative, $R = 0$.

M-4.1.6.2.4. Using Table B-24 of Appendix B, we find the critical value $w_{0.05,10} = 11$.

Table M-4.
TCE Concentrations (mg/L) at Monitoring Wells Before and After Treatment for Example M-4.1.6.2

Sample	Baseline (01/2000)	Post–Test (12/2000)	Difference	Signed Rank
Well 1	20.9	0.917	–20.0	–6
Well 2	9.17	8.77	–0.400	–1
Well 3	5.96	4.37	–1.59	–2
Well 4	41.5	4.34	–37.2	–9
Well 5	34.3	10.7	–23.6	–7
Well 6	19.7	1.48	–18.2	–5
Well 7	38.9	0.272	–38.6	–10
Well 8	8.18	0.520	–7.66	–4
Well 9	9.13	3.06	–6.07	–3
Well 10	28.5	1.90	–26.6	–8

M-4.1.6.2.5. Recall that negative values of the difference support the alternative hypothesis. Therefore we reject H_0 if *R* is smaller than the critical value. Comparing the calculated test statistic and the critical value $R = 0 \leq \{n(n+1)/2\} - w_{\alpha,n} = \{10(11)/2\} - 11 = 44$, so H_0 is rejected. The treatment appears to have lowered TCE concentration in groundwater.

M-4.1.6.2.6. If the differences do not meet the symmetry assumption of the Wilcoxon signed ranks test, the one-sample sign test could be used for the analysis. However, a specific example will not be presented here.

M-4.2. The Quantile Test. The quantile test is used to compare two populations using *m* independent random samples (x_1, x_2,..., x_m) from the first population and *n* independent random samples (y_1, y_2,..., y_n) from the second population. The quantile test is useful in detecting instances where only parts of the data are different rather than a complete shift in the data. It looks at a certain number of the largest data values to determine if too many data values from one population are present to be accounted for by pure chance. When the quantile test and the Wilcoxon rank sum test (discussed above) are applied together, the combined tests are the most powerful at detecting true differences between two populations.

M-4.2.1. Introduction. The quantile test assumes a set of random samples from population 1 and an independent set of random samples from population 2. The quantile test is not robust to outliers, and assumes either a systematic (e.g., a triangular grid) or simple random sampling design. The quantile test may not be used for stratified designs. In addition, exact false rejection error rates are not available, only approximate rates. The quantile test is difficult to do by hand so directions are not included in this guidance. Directions for a modified quantile test that can be done by hand are contained below in Paragraph M-4.2.2, followed by an example in Paragraph M-4.2.3.

M-4.2.2. Directions for a Modified Quantile Test Done by Hand. Let there be *m* measurements from population 1 (the reference area or background group) and *n* measurements from population 2 (the test area). The modified quantile test can be used to detect differences in shape and location of the two distributions. For this test, the significance level, α , can either be approximately 0.10 or approximately 0.05.

M-4.2.2.1 H_0: population 1 = population 2.

M-4.2.2.2. H_A: population 2 > population 1.

M-4.2.2.3. Combine the two samples and orders them from smallest to largest, keeping track of which sample a value came from.

M-4.2.2.4. Using Table B-25 of Appendix B, determine the critical number (*C*) for a sample size *n* from the reference area and sample size *m* from the test area using the significance level α . If the C^{th} largest measurement of the combined population is the same as others, increase *C* to include all of these tied values.

M-4.2.2.4.1. If the largest *C* measurements from the combined samples are all from population 2 (the test area), then reject the null hypothesis and conclude that there are differences between the two populations.

M-4.2.2.4.2. Otherwise, the null hypothesis is not rejected and it appears that there is no difference between the two populations.

M-4.2.3. Example of a Modified Quantile Test Done by Hand. Consider the case where nickel surface soil concentrations are compared between Site A and background using the test (Table M-5).

M-4.2.3.1. H_0: population 1 = population 2.

M-4.2.3.2. H_A: population 1 > population 2.

M-4.2.3.3. Suppose data for nickel surface site data (population 1) are the m = 6 values: 2.67, 3.61, 5.47, 7.15, 8.34, and 7.96 mg/kg.

M-4.2.3.4. Suppose data for nickel surface background data (population 2) are the n = 10 values: 5.14, 7.46, 5.99, 3.36, 3.19, 2.87, 5.95, 1.72, 4.77, and 5.61 mg/kg.

Table M-5.
Nickel Surface Soil Concentrations for Example M-4.2.3

Location	Result (mg/kg)	Rank
Background	1.72	1
Site	2.67	2
Background	2.87	3
Background	3.19	4
Background	3.36	5
Site	3.61	6
Background	4.77	7
Background	5.14	8
Site	5.47	9
Background	5.61	10
Background	5.95	11
Background	5.99	12
Site	7.15	13
Background	7.46	14
Site	7.96	15
Site	8.34	16

M-4.2.3.5. $C_{n,m,\alpha} = C_{10,6,0.05} = 5$; because the fifth largest value is 5.99, there is no need to increase C.

M-4.2.3.6. Only three of the largest five values are from population 1 (site concentrations), therefore the null hypotheses cannot be rejected. The result is that there is no difference between the site concentrations and the background concentrations of nickel.

M-5. Multiple Population Tests. This Paragraph describes procedures to evaluate data from more than two populations. One could accomplish the same objectives by applying the tests described above multiple times. However, doing so would underestimate the true false rejection decision error rate. In other words, if multiple individual tests are done, H_0

is rejected more frequently than desired. The tests described in this Paragraph control the overall false rejection decision error rate by making multiple comparisons simultaneously.

M-5.1. <u>One-Factor Analysis of Variance (ANOVA)</u>. The one-factor ANOVA is a statistical procedure to determine whether differences in mean concentrations among two or more populations are statistically significant. When a single variable is being measured for multiple populations (e.g., the concentration of chromium at multiple sites), the one-factor ANOVA allows the comparison of multiple population means in one test. Because the ANOVA test compares all the means to one another simultaneously, large false positives rates associated with multiple separate pairwise mean comparisons are avoided. Multi-factor ANOVA tests would be used when comparing several variables from multiple populations (e.g., the concentration of arsenic and chromium at multiple sites), but these are more complex than one-factor ANOVA tests and are beyond the scope of this document.

M-5.1.1. <u>Introduction</u>. There are two types of ANOVAs: parametric and nonparametric. The parametric ANOVA assumes that the errors, called residuals, are normally distributed with equal variance. The one-way parametric ANOVA model is the following:

$$x_{i,j} = \mu_i + \varepsilon_{i,j}$$

The $x_{i,j}$ denotes the j^{th} measured value of the i^{th} group, where the i^{th} group contains n_i values and $i = 1, 2, \ldots K$ (the number of groups or populations). The residuals $\varepsilon_{i,j}$ are assumed to be values of a random variable ε that possess a normal distribution with mean of zero and standard deviation of σ. The parameters μ_i are the populations means for the groups; each possessing a common standard deviation σ. The equation is a model in the sense that it is of the form:

Measured value = *Function one or more parameters* + *Residual* (*random error*).

(Also refer to the linear regression model in Appendix Q.) As the population means μ_i are unknown, they are estimated by the sample group means:

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{i,j}}{n_i} \text{ for } i = 1, 2, \ldots K.$$

M-5.1.1.1. Thus, the “true” residuals $\varepsilon_{i,j}$ are estimated by the “sample” residuals as follows:

$$e_{i,j} = x_{i,j} - \bar{x}_i \ .$$

The sample residuals for each group (e.g., the n_i residuals for group i) must each be tested for normality and must be normally distributed.

M-5.1.1.2. The ANOVA is especially useful in situations where sample sizes are small. To apply a parametric one-way ANOVA, at least two groups must be present in the data and at least two samples must be available for each group. Although the ANOVA assumes equal variances, the test is not sensitive to unequal variances as long as the violation is not severe.

M-5.1.1.3. Directions for the ANOVA are given in Paragraph M-5.1.2, followed by an example in Paragraph M-5.1.3.

M-5.1.2. Directions for the ANOVA Test. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the K sample populations to be compared to one another. Let the values from each population be represented by $x_{i,j}$ where $i = 1, 2, \ldots, K$ for the K groups and $j = 1, 2, \ldots, n_i$ for the observations in the i^{th} group.

M-5.1.2.1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ (no difference among the population means).

M-5.1.2.2. H_A: at least one mean, μ_i is different from one or more of the other means.

M-5.1.2.3. Verify that the residuals are normally distributed with equal variances (see Appendix F and Appendix N, respectively).

M-5.1.2.4. Let $(1-\alpha)100\%$ represent the chosen significance level for the test, so α is the false rejection rate for the test. Set up the ANOVA table as follows:

Source of Variation	**Degrees of Freedom (ν)**	**Sum of Squares**	**Mean Square**	***F*-Value**
Groups	$\nu_G = K - 1$	SSG	$\text{MSG} = \text{SSG}/(K-1)$	$F = \dfrac{\text{MSG}}{\text{MSE}}$
Error	$\nu_E = \left(\sum_{i=1}^{K} n_{\text{i}}\right) - 1$	SSE	$\text{MSE} = \text{SSE} \Big/ \left(\sum_{i=1}^{K} n_i - K\right)$	
Total	$\nu_T = \left(\sum_{i=1}^{K} n_{\text{i}}\right) - 1$	SST	$\text{MST} = \text{SST} \Big/ \left(\sum_{i=1}^{K} n_i\right) - 1$	

$$\text{SST} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{i,j}^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{i,j} \Big/ \sum_{i=1}^{k} n_i$$

$$\text{SSG} = \sum_{i=1}^{K} n_i\left(\bar{x}_i - \bar{x}\right)^2 = \sum_{i=1}^{k}\left[\left(\sum_{j=1}^{n_i} x_{i,j}\right)^2 \Big/ n_i\right] - \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{i,j} \Big/ \sum_{i=1}^{k} n_i$$

$$\text{SSE} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}_i\right)^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(e_{i,j}\right)^2 = \text{SST} - \text{SSG} .$$

Note that

$$\nu_T = \nu_G + \nu_E$$

$$\text{SST} = \text{SSG} + \text{SSE} .$$

M-5.1.2.5. It may be convenient to calculate MSE using the formula:

$$\text{MSE} = \frac{\sum_{i=1}^{K}(n_i - 1)\, s_i^2}{\sum_{i=1}^{K} n_i - K} .$$

In this form, MSE is often referred to as the “pooled” variance for the *K* groups, where s_i^2 is the sample variance for the *i*th group:

$$\text{s}_i^2 = \frac{\sum_{j=1}^{n_i}(x_{i,j} - \bar{x}_i)^2}{n_i - 1} .$$

M-5.1.2.6. Use Table B-7 of Appendix B to determine the critical value, $F_{1-\alpha,\nu_G,\nu_E}$, where $F_{\gamma,m,n}$ is the $\gamma 100^{\text{th}}$ percentile of the *F* distribution with *m* degrees of freedom for the numerator and *n* degrees of freedom for the denominator. Compare *F* to $F_{1-\alpha,\nu_G,\nu_E}$. If $F > F_{1-\alpha,\nu_G,\nu_E}$, then reject H_0 (the means of the sample populations are not all equal). Otherwise, conclude that there is no difference among the sample population means. If H_0 is rejected, perform multiple comparison tests to determine which populations are significantly different.

M-5.1.2.7. Statistical software sometimes outputs the coefficient of determination for the ANOVA:

$$r_{ANOVA}^2 = \text{SSG/SST} .$$

The square root of this quantity is similar in function to the regression coefficient for an ordinary least squares regression line (refer to Appendix Q) in that it accounts for the variation in the measured values accounted for by the model (often referred to as the explained variation). A large value for r_{ANOVA}^2 (which ranges from 0 to 1) indicates that most of the variation is ascribable to differences between the group means. It can be shown that

$$F = \frac{r_{ANOVA}^2}{1 - r_{ANOVA}^2} \times \left(\frac{\nu_E}{\nu_G} \right)$$

$$r_{ANOVA}^2 = \frac{\nu_G \, F}{\left(\nu_E + \nu_G F \right)} \ .$$

Therefore, when the calculated value of the statistic F is small (i.e., when the null hypothesis is not rejected), r_{ANOVA}^2 will be near zero.

M-5.1.3. Example of ANOVA. Suppose manganese (Mn) groundwater concentrations are going to be compared among the seven different wells at Site A using the following test with 95% level of confidence.

M-5.1.3.1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ (no difference among the sample means).

M-5.1.3.2. H_A: at least one mean, μ_i is different from one or more of the other means.

M-5.1.3.3. Table M-6 presents the data. All Mn concentrations were detected, so no proxy concentrations are needed to evaluate the data.

M-5.1.3.4. The data were tested for equal variances using Bartlett's test for equal variances (see Paragraph N-3). The data were also tested for normality using the Shapiro-Wilk test. Because the data were not normal, the data were transformed so that the residuals would follow a normal distribution.

M-5.1.3.5. Summary statistics for each well are presented in Table M-7.

M-5.1.3.6. Let $(1-\alpha)100\%$ represent the chosen significance level for the test, where $\alpha = 0.05$. Note that in this example K = 7 and n_i = 8 for i = 1, 2, … 7. Set up the ANOVA table as follows:

Source of Variation	**Degrees of Freedom (ν)**	**Sum of Squares**	**Mean Square**	***F* Value**
Groups	6	137.29	22.88	346.09
Error	49	3.24	0.066	
Total	55	140.53		

M-5.1.3.7. The power of an ANOVA F-test can be estimated prior to a study. Table B-28 in Appendix B lists the power for K = 3 to 10 groups and significance levels of α =

0.2, 0.1, and 0.05, where each group contains an equal number of samples *n*. To use the tables, the "effect size," Δ, must be also estimated as:

$$\Delta = \text{(largest group mean – smallest group mean)/(MSE)}^{1/2} .$$

M-5.1.3.8. The tables list various values of Δ. For a specified value of *K*, *n*, α, and Δ, the tables list the minimum power (probability) corresponding to the alternative hypothesis that all group means, other than the two extremes, are equal to the "grand mean," which is equal to the median of the largest and smallest group means. When comparing *K* groups of equal size *n*, the tables are useful for determining approximately how large a sample size for each group is required to achieve a particular level of confidence 1 – α and power 1 – β. For example, for $K = 3$ groups and $\alpha = 0.05$, to detect a size effect $\Delta = 1.0$ (i.e., a difference between the largest and smallest mean equal to $\text{MSE}^{1/2}$) with power of at least $1 - \beta = 0.80$, the required sample size for each group $n \approx 20$.

Table M-6.
Manganese (Mn) Groundwater Concentrations to be Compared Among the Wells at Site A

Well Location	Result	Log Result mg/L	Well Location	Result mg/L	Log Result
69-2-02	0.432	–0.839	69-2-06A	0.294	–1.224
69-2-02	0.44	–0.821	69-2-06A	0.301	–1.201
69-2-02	0.513	–0.667	69-2-06A	0.379	–0.970
69-2-02	0.704	–0.351	69-2-06A	0.352	–1.044
69-2-02	0.327	–1.118	69-2-06A	0.346	–1.061
69-2-02	0.316	–1.152	69-2-06B	0.13	–2.040
69-2-02	0.454	–0.790	69-2-06B	0.184	–1.693
69-2-02	0.401	–0.914	69-2-06B	0.209	–1.565
69-2-04	0.0504	–2.988	69-2-06B	0.2	–1.609
69-2-04	0.0502	–2.992	69-2-06B	0.0739	–2.605
69-2-04	0.054	–2.919	69-2-06B	0.0876	–2.435
69-2-04	0.0523	–2.951	69-2-06B	0.126	–2.071
69-2-04	0.0923	–2.383	69-2-06B	0.129	–2.048
69-2-04	0.0556	–2.890	69-2-07	0.0137	–4.290
69-2-04	0.0534	–2.930	69-2-07	0.019	–3.963
69-2-04	0.0517	–2.962	69-2-07	0.0163	–4.117
69-2-05	0.00684	–4.985	69-2-07	0.0195	–3.937
69-2-05	0.00639	–5.053	69-2-07	0.0112	–4.492
69-2-05	0.00631	–5.066	69-2-07	0.0112	–4.492
69-2-05	0.00813	–4.812	69-2-07	0.0102	–4.585
69-2-05	0.00747	–4.897	69-2-07	0.00946	–4.661
69-2-05	0.00679	–4.992	69-2-08	0.563	–0.574
69-2-05	0.00731	–4.919	69-2-08	0.512	–0.669

Well Location	Result	Log Result mg/L	Well Location	Result mg/L	Log Result
69-2-05	0.00444	–5.417	69-2-08	0.475	–0.744
69-2-06A	0.3	–1.204	69-2-08	0.546	–0.605
69-2-06A	0.286	–1.252	69-2-08	0.276	–1.287
69-2-06A	0.303	–1.194	69-2-08	0.383	–0.960
			69-2-08	0.33	–1.109
			69-2-08	0.27	–1.309

Table M-7.
Summary Statistics for Mn by Well

Well	Sample Size	Mean of Log Result	Standard Deviation of Log Result
69-2-02	8	–0.832	0.2539
69-2-04	8	–2.877	0.2026
69-2-05	8	–5.018	0.1818
69-2-06A	8	–1.144	0.1031
69-2-06B	8	–2.008	0.3779
69-2-07	8	–4.317	0.2832
69-2-08	8	–0.907	0.3011

M-5.2. Kruskal-Wallis Test. The Kruskal-Wallis test is the nonparametric version of the ANOVA. It is a statistical procedure to determine whether differences in median concentrations among a number of groups or multiple populations are statistically significant. The Kruskal-Wallis allows the comparison of multiple population means in one test. If the test shows statistically significant differences among the groups, multiple comparison procedures can be used to identify which group or groups are different.

M-5.2.1. Introduction. In terms of hypothesis tests, the null hypothesis is that all group medians are equal and the alternative hypothesis is that at least one group is different from one or more other groups. To test this hypothesis, no assumptions are required about the shape of the distributions; each group may have a different distribution. The Kruskal-Wallis test is used to evaluate whether the distributions are identical. Directions for the Kruskal-Wallis test are given below in Paragraph M-5.2.2, followed by an example in Paragraph M-5.2.3.

M-5.2.2. Directions for the Kruskal-Wallis Test. Let $(1-\alpha)100\%$ represent the chosen significance level for the test.

M-5.2.2.1. Rank all $x_{i,j}$ observations from lowest to highest. Let $R_{I,j}$ denote the rank of the $x_{i,j}$ observation.

M-5.2.2.1.1. Ties. If two or more observations are numerically equal, then use an average rank for each observation. The average rank is calculated as the average of the ranks that the tied observations would have received had the observations been different.

M-5.2.2.1.2. Censored Data. If any values are not-detected, it is appropriate to consider the ranks for these values equal to zero. (It is irrelevant what number is assigned to the non-detected values as long as all such values are assigned the same number, and it is smaller than any detected value.)

M-5.2.2.2. Add the ranks of the observations in each group. Call the sum of the ranks for the i^{th} group R_i. Also calculate the average rank for each group, $\overline{R}_i = R_i / n_i$. If there are at least 50% detected results and no tied values, then compute the Kruskal-Wallis statistic:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{K} n_i \left(\overline{R}_i - \frac{N+1}{2} \right)^2$$

$$= \left\lfloor \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} \right\rfloor - 3(N+1)$$

where

$$N = \sum_{i=1}^{K} n_i \ .$$

M-5.2.2.3. If there are at least 50% detected results and there are tied values present in the data, then compute the adjusted Kruskal-Wallis statistic:

$$H' = \frac{\left\lfloor \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} \right\rfloor - 3(N+1)}{1 - \left(\sum_{k=1}^{g} W_k \Big/ (N^3 - N) \right)}$$

where g is the number of groups of distinct tied observations and $W_k = (w_k^3 - w_k)$, where w_k is the number of tied data in the tied group k. Note that the unique observations can be considered groups of size 1, with the corresponding $W_k = (1^3 - 1) = 0$. If all the group medians are equal, then $H = 0$. As the differences between the groups medians increase, H will also increase; so the larger the value of H, the less probable H_0 is true.

M-5.2.2.4. Compare the calculated value H (or H') to the tabulated critical value for the chi-square distribution, $\chi^2_{1-\alpha,K-1}$, with $K - 1$ degrees of freedom and $(1-\alpha)100\%$ level of confidence (found in Table B-2 of Appendix B).

M-5.2.2.5. Reject H_0 if $H > \chi^2_{1-\alpha,K-1}$. If H_0 is rejected use multiple comparison tests to determine which populations are significantly different.

M-5.2.3. Example of the Kruskal-Wallis Test. Suppose lead groundwater concentrations are going to be compared among seven wells using the Kruskal-Wallis test with 95% level of confidence.

M-5.2.3.1. $H_0 : \mu_1 = \mu_2 = \ldots = \mu_7$ (i.e., no difference among the well means).

M-5.2.3.2. $H_A :$ at least one mean is different from one or more of the other means.

M-5.2.3.3. Table M-8 presents the data. All lead concentrations were detected so no proxy concentrations were needed to evaluate the data.

M-5.2.3.4. The sum of the ranks for each of the seven groups is:

$$R_1 = 272,\ R_2 = 168,\ R_3 = 62.5,\ R_4 = 420,\ R_5 = 304,\ R_6 = 73.5,\ R_7 = 296$$

M-5.2.3.5. Because there are at least 50% detected results and there are tied values present in the data, compute the adjusted Kruskal-Wallis statistic:

$$H' = \frac{\left\lfloor \dfrac{12}{N(N+1)} \displaystyle\sum_{i=1}^{K} \frac{R_i^2}{n_i} \right\rfloor - 3(N+1)}{1 - \left(\displaystyle\sum_{k=1}^{g} W_k \big/ (N^3 - N) \right)}$$

The table below summarizes the g = 4 tied groups:

Tied Rank	Number of Tied Observations w_k	$W_K = w_k^3 - w_k$
4	3	24
12.5	2	6
19.5	2	6
21.5	2	6

$$H' = \frac{\left\lfloor \dfrac{12}{56(56+1)} \times \left(\dfrac{272^2}{8} + \dfrac{168^2}{8} + \dfrac{62.5^2}{8} + \dfrac{420^2}{8} + \dfrac{304^2}{8} + \dfrac{73.5^2}{8} + \dfrac{296^2}{8} \right) \right\rfloor - 3(56+1)}{1 - \left[(24+6+6+6)/(56^3 - 56) \right]} = 48.91$$

$$\chi^2_{1-\alpha, K-1} = \chi^2_{1-0.05, 7-1} = \chi^2_{0.95, 6} = 12.59.$$

M-5.2.3.6. Now compare the calculated value to the critical value, 48.91 > 12.59. As the calculated value exceeds the critical value, reject $H_{0.}$

M-5.2.3.7. Because there is a difference in the average lead concentration among the seven wells, a multiple comparison test should be done to determine which wells are significantly different. A multiple comparison test based on ranks is discussed in Conover (1980).

Table M-8.
Lead Concentrations for Example M-5.2.3

Well	Result mg/L	Rank	Well	Result mg/l	Rank
6	0.978	1	5	3.100	29
6	1.037	2	7	3.118	30
3	1.061	4	5	3.144	31
3	1.061	4	7	3.178	32
3	1.061	4	1	3.215	33
6	1.095	6	1	3.219	34
6	1.109	7	1	3.235	35
3	1.144	8	5	3.346	36
3	1.227	9	1	3.395	37
3	1.241	10	5	3.421	38
3	1.270	11	5	3.434	39
3	1.426	12.5	1	3.478	40
6	1.426	12.5	1	3.586	41
6	1.513	14	5	3.605	42
6	1.530	15	5	3.627	43
6	1.601	16	7	3.671	44
2	2.588	17	7	3.689	45
2	2.595	18	5	3.694	46
2	2.610	19.5	7	3.922	47
2	2.610	19.5	7	3.932	48
2	2.625	21.5	4	4.057	49
2	2.625	21.5	4	4.101	50
2	2.639	23	4	4.103	51
7	2.918	24	4	4.119	52
1	3.011	25	4	4.159	53
7	3.035	26	4	4.177	54
1	3.068	27	4	4.214	55
2	3.073	28	4	4.228	56

M-6. Multiple Comparison Tests. Multiple comparisons occur whenever more than one statistical test is performed with the same data. These comparisons can arise, for example, as a result of the need to test multiple down-gradient wells against a pool of up-gradient background data or to regularly test several indicator parameters for contamination. The multiple comparison tests described in this section may not be needed if a significant difference is not obtained from the ANOVA *F*-test.

M-6.1. Introduction. Comparisons are usually written in terms of linear combinations of the population means, and are often referred to as “contrasts.” For example, we may want to know if the mean for population 1, μ_1, differs from the mean for population 2, μ_2. This contrast can be written as $\mu_1 - \mu_2$. In general, a contrast is a linear combination

$$\theta = \sum a_i \mu_i$$

where

$$\sum a_i = 0.$$

Beyond comparing pairs of means, a contrast to compare the mean of population 1 to the means of populations 2 and 3 can be written as $2\mu_1 - \mu_2 - \mu_3$.

M-6.1.1. The Type I error rate for multiple comparison tests can be viewed in two ways. Comparison-wise significance considers the probability of rejecting the hypothesis that only a single contrast equals zero ($H_0 : \theta_1 = 0$) when it is actually true. Experiment-wise significance considers the probability of rejecting any of a set of *m* hypotheses on contrasts ($H_0 : \theta_j = 0, j = 1, ..., m$) when all of them are true.

M-6.1.2. Table M-9 summarizes the multiple comparison tests that will be covered in this document. The Fisher’s Least Significant Difference (LSD) test and Bonferroni’s test are multiple comparison tests that are based on the Student’s *t* distribution, whereas the Tukey’s test and Duncan’s multiple range test are based on the Studentized range statistic. Scheffé’s multiple comparison test is used to achieve an experiment-wise false positive rate for all possible contrasts or linear combinations of means at the same time. All the multiple comparison tests presented rely on the assumption of normality. Assumptions of normality should have been verified during the ANOVA process, which is typically performed prior to these multiple comparison tests. More information on multiple comparison tests can be found in Mason et al. (1989) and Montgomery (1997).

M-6.1.3. There is no clear answer to the question of which multiple comparison technique should be used. For comparing all pairs of treatment means, Fisher’s LSD is the least conservative (most powerful) test for identifying differences between means (i.e., it rejects H_0 most often) followed by Duncan’s Multiple Range, Tukey, and finally Sheffé. The relative conservatism of the Bonferroni Test will depend on the number of groups. Montgomery (1997) recommends Fisher’s LSD or Duncan’s multiple range test for comparing all treatment means as long as the ANOVA *F*-test is significant, based on Monte Carlo studies conducted by Carmer and Swanson (1973). Mason et al. (1989) recommend Fisher’s LSD to control the comparison-wise error rate and Tukey’s test to control the experiment-wise error rate for comparing all treatment means. When many comparisons need to be made, multiple range tests such as Duncan's multiple range test and Tukey's test should be used as a compromise between the desired experiment-wise error rate and an unacceptable comparison-wise error rate (Mason et al., 1989). Obviously, if one's purpose is to compare treatment means to a control or if contrasts other than pairwise comparisons of treatments are of interest, Dunnett's, Bonferroni's, or Scheffé's test may be preferred.

Table M-9.
Summary of Multiple Comparison Tests

Test	Purpose
Dunnett's	Comparing treatment means to a control mean
Fisher's LSD	Comparing all pairs of means
Duncan's multiple range	Comparing all pairs of means
Tukey's	Comparing all pairs of means
Bonferroni's	Comparing any set of contrasts
Scheffé's	Comparing any set of contrasts

M-6.2. Fisher's Least-Significant Difference Test. Fisher's LSD test is an extension of the *t*-test for comparing all pairs of treatment means. Each pairwise comparison will have a Type I error rate (probability of declaring the pair of means different when they are not) of α. Therefore, the *experiment-wise* error rate (the probability of declaring any pair of means different when they are not) will be *larger* than α. The disadvantage to the Fisher's LSD test is that its experiment-wise error rate is not satisfactory for testing all possible pairs of group means when there are a moderate to large number of groups to be compared (Mason et al., 1989). Directions for Fisher's LSD test (from Mason et al., 1989) are given in Paragraph M-6.2.1 and an example is presented in Paragraph M-6.2.2.

M-6.2.1. Directions for Fisher's LSD Test. Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let the values from each population be represented by $x_{i,j}$ where *i* = 1, 2,…, *K* for the *K* groups and $j = 1, 2, ..., n_i$ for the observations at the i^{th} group. Let $(1-\alpha)100\%$ represent the chosen confidence level for the test.

M-6.2.1.1. Verify the assumptions of normality.

M-6.2.1.2. The means of two groups, $\bar{x}_i$ and $\bar{x}_k$, in an ANOVA are declared to be significantly different if:

$$\bar{x}_i - \bar{x}_k > \text{LSD}$$

where

$$\text{LSD} = t_{1-\alpha/2,v_E}\left[\text{MSE}\left(\frac{1}{n_i}+\frac{1}{n_k}\right)\right]^{1/2}$$

and

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} \ .$$

t_{γ,v_E} is the γ100th percentile for the Student's *t* distribution with v_E degrees of freedom (see Table B-23 in Appendix B). MSE and v_E come from the ANOVA procedures previously defined. Note that for *K* groups, $K(K-1)/2$ differences $\bar{x}_i - \bar{x}_k$ need to be calculated.

M-6.2.2. Example of Fisher's LSD Test. Mean manganese groundwater concentrations in seven wells were compared to one another using the ANOVA. The null hypothesis was rejected. The LSD test is subsequently applied below using the 95% level of confidence.

M-6.2.2.1. The table in Paragraph M-5.1.3 presents the data. All manganese concentrations were detected so no proxy concentrations are needed to evaluate the data.

M-6.2.2.2. Assumptions of normality were verified for the log result during the ANOVA process.

$$\begin{aligned}\text{LSD} &= t_{1-\alpha/2,v_E}\left[\text{MSE}\left(\frac{1}{n_i}+\frac{1}{n_k}\right)\right]^{1/2}\\ &= t_{0.975,49}\left[0.066\times\left(\frac{1}{8}+\frac{1}{8}\right)\right]^{1/2}\\ &= 2.01\times 0.128\\ &= 0.2584 \ .\end{aligned}$$

M-6.2.2.3. Means that differ by more than 0.2584 would be considered statistically different with 95% confidence. Alternatively, confidence intervals for the difference in means can be calculated as $(\bar{x}_i - \bar{x}_k) \pm \text{LSD}$. If zero is not in the confidence interval, the two population means are declared significantly different at the α significance level. Table M-10 summarizes the results. Comparisons significant at the 0.05 level are indicated by ***.

M-6.2.2.4. Another way to visualize the conclusions is to list the means in order and identify those that are not significantly different. In Table M-11, means designated with the same "group" letter (A, B, C, etc.) are not significantly different at $\alpha = 0.05$.

M-6.2.2.5. As Wells 69-2-02 and 69-2-08 are in LSD grouping A, the means for these wells are not statistically different. The preceding table indicates that the difference between the two means is 0.0758, which is less than LSD = 0.2584.

M-6.3. Bonferroni's Test. The Bonferroni's test is designed to control the *experiment-wise* error rate (the probability of declaring any two means different when they are not). The test uses the overall significance level divided by the number of selected comparisons as the comparison-wise significance level. Mason et al. (1989) warn that Bonferroni's test should not be used when the number of comparisons becomes very large, because this results in an extremely conservative comparison-wise test. However, they do state that the experiment-wise error rate can be better controlled using Bonferroni's test rather than the Fisher's LSD test (where comparison-wise error is controlled). Also, note that Bonferroni's test can be used to test any contrast of interest (Mason et al., 1989). Directions for Bonferroni's Test (from Mason et al., 1989) are presented in Paragraph M-6.3.1 and an example is presented in Paragraph M-6.3.2.

Table M-10.
Results for Example M-6.2.2

Well Comparison	Difference Between Means $\bar{x}_i - \bar{x}_k$ (mg/L)	95% Confidence Interval (mg/L)
02 – 08	0.0758	(–0.1825, 0.3342)
02 – 06A	0.3123	(0.0539, 0.5706)***
02 – 06B	1.1769	(0.9186, 1.4353)***
02 – 04	2.0452	(1.7868, 2.3036)***
02 – 07	3.4857	(3.2273, 3.7440)***
02 – 05	4.1861	(3.9277, 4.4444)***
08 – 06A	0.2365	(–0.0219, 0.4948)
08 – 06B	1.1011	(0.8427, 1.3595)***
08 – 04	1.9694	(1.7110, 2.2277) ***
08 – 07	3.4098	(3.1515, 3.6682)***
08 – 05	4.1103	(3.8519, 4.3686)***
06A – 06B	0.8646	(0.6063, 1.1230)***
06A – 04	1.7329	(1.4746, 1.9913)***
06A – 07	3.1734	(2.9150, 3.4317)***
06A – 05	3.8738	(3.6154, 4.1322)***
06B – 04	0.8683	(0.6099, 1.1266)***
06B – 07	2.3088	(2.0504, 2.5671)***
06B – 05	3.0092	(2.7508, 3.2675)***
04 – 07	1.4405	(1.1821, 1.6988)***
04 – 05	2.1409	(1.8825, 2.3992)***
07 – 05	0.7004	(0.4420, 0.9588)***

Table M-11.
List of the Means in Order for Example M-6.2.2

Well	Mean of Log Result	*n*	LSD Groupings
69-2-02	–0.8315	8	A
69-2-08	–0.9073	8	B A
69-2-06A	–1.1438	8	B
69-2-06B	–2.0084	8	C
69-2-04	–2.8767	8	D
69-2-07	–4.3172	8	E
69-2-05	–5.0176	8	F

M-6.3.1. Directions for Bonferroni's Test. Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let the values from each population be represented by $x_{i,j}$ where *i* = 1, 2,…, *K* for the *K* groups and $j = 1, 2, ..., n_i$ for the observations in the *i*th group. Let $(1-\alpha)100\%$ represent the selected confidence level for the test.

M-6.3.1.1. Verify the assumptions of normality.

M-6.3.1.2. Let

$$\theta = \sum a_i \mu_i$$

represent one of *m* linear combinations of the means, μ_k, for which the hypothesis $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$ is being tested.

M-6.3.1.3. Reject H_0 if

$$|\theta| = \left| \sum a_i \bar{x}_i \right|$$

exceeds

$$\text{BSD} = t_{1-\alpha/2m, \nu_E} \left[\text{MSE} \sum a_i^2 / n_i \right]^{1/2}$$

where n_i is the number of observations used to calculate

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j}$$

.

t_{γ,v_E} is the $\gamma 100^{\text{th}}$ percentile for the Student's *t* distribution with v_E degrees of freedom (see Table B-23 in Appendix B), and *m* is the number of comparisons. For *K* means (groups), there are

$$m = \frac{K(K-1)}{2}$$

possible comparisons. MSE and v_E are determined from the ANOVA procedures previously defined.

M-6.3.2. <u>Example of Bonferroni's Test</u>. Suppose manganese groundwater concentrations are going to be compared among the seven different wells at Site A using Bonferroni's test with 95% level of confidence.

M-6.3.2.1. Table M-6 presents the data. All manganese concentrations were detected, so no proxy concentrations are needed to evaluate the data.

M-6.3.2.2. The assumptions of normality were verified during the ANOVA process. The contrasts to make pairwise comparisons of all 7 means are the 21 differences (where $a_i = \pm 1$):

$$\mu_{69-2-02} - \mu_{69-2-04}$$
$$\mu_{69-2-02} - \mu_{69-2-05}$$
$$\mu_{69-2-02} - \mu_{69-2-06A}$$
$$\mu_{69-2-02} - \mu_{69-2-06B}$$
$$\mu_{69-2-02} - \mu_{69-2-07}$$
$$\mu_{69-2-02} - \mu_{69-2-08}$$
$$\mu_{69-2-04} - \mu_{69-2-05}$$
$$\mu_{69-2-04} - \mu_{69-2-06A}$$
$$\mu_{69-2-04} - \mu_{69-2-06B}$$
$$\mu_{69-2-04} - \mu_{69-2-07}$$
$$\mu_{69-2-04} - \mu_{69-2-08}$$
$$\mu_{69-2-05} - \mu_{69-2-06A}$$
$$\mu_{69-2-05} - \mu_{69-2-06B}$$
$$\mu_{69-2-05} - \mu_{69-2-07}$$
$$\mu_{69-2-05} - \mu_{69-2-08}$$
$$\mu_{69-2-06A} - \mu_{69-2-06B}$$
$$\mu_{69-2-06A} - \mu_{69-2-07}$$
$$\mu_{69-2-06A} - \mu_{69-2-08}$$
$$\mu_{69-2-06B} - \mu_{69-2-07}$$
$$\mu_{69-2-06B} - \mu_{69-2-08}$$
$$\mu_{69-2-07} - \mu_{69-2-08}$$

$$\text{BSD} = t_{1-\alpha/2m, v_E}\left[\text{MSE}\sum a_i^2 / n_i\right]^{1/2} = t_{0.999,49}\left[0.066\left(\frac{1}{8}+\frac{1}{8}\right)\right]^{1/2} = 3.20 \times 0.128 = 0.412 \ .$$

Means that differ by more than 0.412 would be considered statistically different with 95% confidence. Alternatively, confidence intervals for the difference in means can be calculated as $\bar{x}_i - \bar{x}_k \pm \text{BSD}$. If zero is not covered by the confidence interval, the two population means are declared significantly different at the α significance level.

M-6.3.2.3. In Table M-12, means with the same letter are not significantly different at $\alpha = 0.05$. For example, the mean for 69-2-02 does not differ from the mean for 69-2-08 by more than 0.412, so we accept

$$H_0 : \mu_{69-2-02} - \mu_{69-2-08} = 0\,.$$

Table M-12.
Means with the Same Letter are not Significantly Different at $\alpha = 0.05$ in Example M-6.3.2

Well	Mean of Log Result	*n*	Bonferroni Grouping
69-2-02	–0.8315	8	A
69-2-08	–0.9073	8	A
69-2-06A	–1.1438	8	A
69-2-06B	–2.0084	8	B
69-2-04	–2.8767	8	C
69-2-07	–4.3172	8	D
69-2-05	–5.0176	8	E

On the other hand, we can reject

$$H_0 : \mu_{69-2-02} - \mu_{69-2-05} = 0$$

because the two observed means differ by more than 0.412. Notice that the more conservative Bonferroni test does not reject

$$H_0 : \mu_{69-2-02} - \mu_{69-2-06A} = 0$$

with 95% confidence while Fisher’s LSD tests did.

M-6.4. Tukey’s Test. Tukey’s test is designed to control the experiment-wise chance of a Type I error (declaring any two population means different when they are not) at α assuming equal sample sizes (Mason et al., 1989). Because of this, it is less powerful than Fisher’s LSD or Duncan’s multiple range test (Montgomery, 1997). Directions and an example for Tukey’s Test (from Mason et al., 1989) are given in Paragraphs M-6.4.1 and M-6.4.2, respectively.

M-6.4.1. Directions for Tukey’s Test. Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let the values from each population be represented by $x_{i,\,j}$, where *i* = 1, 2,…, *K* for the *K* groups and $j = 1, 2, ..., n_i$ for the observations at the *i*th group. Let $(1-\alpha)100\%$ be the confidence level.

M-6.4.1.1. Verify the assumptions of normality. Two averages, $\bar{x}_i$ and $\bar{x}_r$, are based on n_i and n_r samples, respectively, where

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} \ .$$

Two means are significantly different if $|\,\bar{x}_i - \bar{x}_r\,| > \mathrm{TSD}$ where:

$$TSD = q_{\alpha,k,v_E}\left[MSE\left(\frac{1/n_i + 1/n_r}{2}\right)\right]^{1/2} \ .$$

M-6.4.1.2. The quantity q_{α,k,v_E} is the Studentized range statistic in Table B-22 of Appendix B, where *k* is the number of means being compared (typically equal to the number of groups *K*); MSE and v_E are from the ANOVA procedure previously defined, and α represents the desired significance level.

M-6.4.2. Example of Tukey's Test. Manganese groundwater concentrations are compared among the seven different wells at Site A using Tukey's Test with 95% level of confidence.

M-6.4.2.1. Table M-6 presents the data. All manganese concentrations were detected so no proxy concentrations are needed to evaluate the data. Assumptions of normality were verified during the ANOVA process.

$$\begin{aligned} TSD &= q_{\alpha,k,v_E}\left[MSE\left(\frac{1/n_i + 1/n_r}{2}\right)\right]^{1/2} \\ &= q_{0.05,7,49}\left[0.066\left(\frac{1/8+1/8}{2}\right)\right]^{1/2} = 4.35\times 0.0908 = 0.3952 \ . \end{aligned}$$

Table M-13.
Results from Example M-6.4.2

Well Comparison	Difference Between Means (mg/L)	Simultaneous 95% Confidence Intervals (mg/L)
69-2-02–69-2-08	0.0758	(–0.3194, 0.4710)
69-2-02–69-2-06A	0.3123	(–0.0829, 0.7075
69-2-02–69-2-06B	1.1769	(0.7817, 1.5721)***
69-2-02–69-2-04	2.0452	(1.6500, 2.4404) ***
69-2-02–69-2-07	3.4857	(3.0905, 3.8809)***
69-2-02–69-2-05	4.1861	(3.7909, 4.5813)***
69-2-08–69-2-06A	0.2365	(–0.1587, 0.6317)
69-2-08–69-2-06B	1.1011	(0.7059, 1.4963)***
69-2-08–69-2-04	1.9694	(1.5742, 2.3646)***
69-2-08–69-2-07	3.4098	(3.0146, 3.8051)***

Well Comparison	Difference Between Means (mg/L)	Simultaneous 95% Confidence Intervals (mg/L)
69-2-08–69-2-05	4.1103	(3.7150, 4.5055)***
69-2-06A–69-2-06B	0.8646	(0.4694, 1.2598)***
69-2-06A–69-2-04	1.7329	(1.3377, 2.1281)***
69-2-06A–69-2-07	3.1734	(2.7782, 3.5686)***
69-2-06A–69-2-05	3.8738	(3.4786, 4.2690)***
69-2-6B–69-2-04	0.8683	(0.4731, 1.2635)***
69-2-06B–69-2-07	2.3088	(1.9135, 2.7040)***
69-2-06B–69-2-05	3.0092	(2.6139, 3.4044)***
69-2-04–69-2-07	1.4405	(1.0453, 1.8357)***
69-2-04–69-2-05	2.1409	(1.7457, 2.5361)***
69-2-07–69-2-05	0.7004	(0.3052, 1.0956)***

M-6.4.2.2. Means that differ by more than 0.3952 would be considered statistically different with 95% confidence. Alternatively, confidence intervals for the difference in means can be calculated for the difference of any two means as $\bar{x}_i - \bar{x}_r \pm \text{TSD}$. If zero is not in the confidence interval, the two population means are significantly different at the α significance level. Table M-13 summarizes the results. Comparisons significant at $\alpha = 0.05$ are indicated by ***.

M-6.4.2.3. In Table M-14, means with the same letter are not significantly different at $\alpha = 0.05$.

Table M-14.
Means with the Same Letter are not Significantly Different at $\alpha = 0.05$

Tukey Grouping	Mean of Log Result	*n*	Well
A	–0.8315	8	69-2-02
A	–0.9073	8	69-2-08
A	–1.1438	8	69-2-06A
B	–2.0084	8	69-2-06B
C	–2.8767	8	69-2-04
D	–4.3172	8	69-2-07
E	–5.0176	8	69-2-05

M-6.5. <u>Duncan's Multiple Range Test</u>. Duncan's multiple range test is used to test for differences in all pairs of means. Considering the ordered list of means, this procedure provides an experiment-wise error rate of

$$1-(1-\alpha)^{p-1}$$

when the pair of means are *p* steps apart in the ordered list (Montgomery, 1997). Thus, the experiment-wise probability of a Type I error depends on how far apart in the ordered list the two means lie (Mason et al., 1989). Duncan's multiple range test is similar to Tukey's test except that it has greater power to detect differences but does not control the experiment-wise error rate as well. Directions for Duncan's multiple range test (from Mason et al., 1989, and Montgomery, 1997) are presented in Paragraph M-6.5.1 followed by an example in Paragraph M-6.5.2.

M-6.5.1. Directions for Duncan's Multiple Range Test. Let *K* represent the total number of populations to be compared. Let n represent the sample sizes of each of the *K* sample populations. Let the values from each population be represented by $x_{i,j}$ where *i* = 1, 2,…, *K* and *j* = 1, 2,…, *n* for the observations in the *i*th group (population).

M-6.5.1.1. Verify the assumptions of normality. The means

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n_i} x_{i,j}$$

are sorted from smallest to largest. The two extreme means are compared first. The largest and smallest of $p = K$ averages, $\bar{x}_a$ and $\bar{x}_b$ (each based on a sample size of *n*), are significantly different if $\left|\bar{x}_a - \bar{x}_b\right| > R_p$ where

$$R_p = q_{\alpha,p,v_E}\left(\frac{\text{MSE}}{n}\right)^{1/2}.$$

M-6.5.1.2. The quantity

$$q_{\alpha,p,v_E}$$

is the Studentized range critical value (see Table B-6 of Appendix B). MSE and v_E are from the ANOVA procedure previously defined, and α represents the comparison-wise error rate. The experiment-wise significance level for comparing the extremes of *p* means is

$$\alpha_p = 1 - (1-\alpha)^{p-1}.$$

M-6.5.1.2.1. If the smallest and largest means are not significantly different, then no more comparisons are made and all other comparisons are declared not significantly different at the $(1-\alpha_p)100\%$ level of confidence.

M-6.5.1.2.2. If the smallest and largest averages are significantly different, then two comparisons are made where $p = k - 1$: one for the second smallest and the largest averages, and one for the smallest and the second largest averages.

M-6.5.1.2.3. For the two comparisons, if neither of these tests is significantly different, then no more comparisons are performed and only two extreme means ($\bar{x}_a$ and $\bar{x}_b$) are concluded to be significantly different.

M-6.5.1.2.4. If one or both of these tests are statistically significant, testing should continue with groups of averages lying within the two extremes that have been declared significantly different.

M-6.5.1.3. Testing continues until no further significant differences are obtained.

M-6.5.2. Example of Duncan's Multiple Range Test. Suppose manganese groundwater concentrations are going to be compared among the seven different wells at Site A using Duncan's multiple range test with 95% level of confidence.

M-6.5.2.1. Table M-6 presents the data. All manganese concentrations were detected so no proxy concentrations are needed to evaluate the data.

M-6.5.2.2. The assumptions of normality were verified during the ANOVA process.

M-6.5.2.3. There are seven groups to compare so we begin by comparing the one with the smallest mean to the one with the largest mean.

$$R_7 = q_{\alpha,7,v_E}\left(\mathrm{MSE}/n\right)^{1/2} = q_{0.05,7,49}\left(0.066/8\right)^{1/2} = 3.255 \times 0.0908 = 0.296.$$

Considering

$$\left|\,\bar{x}_{69-2-02} - \bar{x}_{69-2-05}\right| = \left|-0.8315 - (-5.0176)\right| = 4.186 > 0.296$$

we can conclude that the population means for these two wells differ at the

$$1-(1-\alpha)^{p-1} = 1-(1-0.05)^{7-1} = 0.26$$

significance level. As the two extreme means were significantly different, we now test means that are 6 levels apart.

$$R_6 = q_{\alpha,6,v_E}\left(\mathrm{MSE}/n\right)^{1/2} = q_{0.05,6,49}\left(0.066/8\right)^{1/2} = 3.212 \times 0.0908 = 0.292.$$

Considering

$$|\,\bar{x}_{69-2-02} - \bar{x}_{69-2-07}| = |-0.8315 - (-4.3172)| = 3.486 > 0.292$$

and

$$|\,\bar{x}_{69-2-08} - \bar{x}_{69-2-05}| = |-0.9073 - (-5.0176)| = 4.110 > 0.292$$

we can conclude that the population means for these two comparisons differ at the

$$1-(1-\alpha)^{p-1} = 1-(1-0.05)^{6-1} = 0.23$$

significance level.

M-6.5.2.4. Because means 6 levels apart are significantly different, continue the process with means 5 levels apart and so on. The final results are summarized in the Table M-15, where means with the same letter are not significantly different at an experiment-wise significance level of α = 0.05.

M-6.6. Dunnett's Test for Simple Random and Systematic Samples. Dunnett's test is used to test the difference between sample or "treatment" means from different populations against a control population. Dunnett's method is the same as the standard two-sample *t*-test (Paragraph M-2), except for the use of a larger pooled estimate of variance and the need for special *t* type tables (Table B-26 of Appendix B). The experiment-wise significance level for all comparisons will be α (Montgomery, 1997). Directions for the use of Dunnett's method for a simple random sample or a systematic random sample are presented in Paragraph M-6.6.1 and followed by an example in Paragraph M-6.6.2.

Table M-15.
Means with the same Letter are not Significantly Different at Significance of α = 0.05

Duncan Grouping	Mean of Log Result	*N*	Well
69-2-02	–0.8315	8	A
69-2-08	–0.9073	8	B A
69-2-06A	–1.1438	8	B
69-2-06B	–2.0084	8	C
69-2-04	–2.8767	8	D
69-2-07	–4.3172	8	E
69-2-05	–5.0176	8	F

M-6.6. Dunnett's Test for Simple Random and Systematic Samples. Dunnett's test is used to test the difference between sample or "treatment" means from different populations against a control population. Dunnett's method is the same as the standard two-sample *t*-test (Paragraph M-2), except for the use of a larger pooled estimate of variance and the need for special *t* type tables (Table B-26 of Appendix B). The experiment-wise significance level for all comparisons will be α (Montgomery, 1997). Directions for the use of

Dunnett's method for a simple random sample or a systematic random sample are presented in Paragraph M-6.6.1 and followed by an example in Paragraph M-6.6.2.

M-6.6.1. Directions for Dunnett's Test for Simple Random and Systematic Samples. Let K represent the total number of populations to be compared so there are (K – 1) sample populations and a single control population. Let $n_1, n_2, \ldots, n_{K-1}$ represent the sample sizes of each of the (K – 1) sample populations and let m represent the sample size of the control population.

M-6.6.1.1. H_0: $\mu_i - \mu_C \le 0$ (no difference between the sample means and the control mean).

M-6.6.1.2. H_A: $\mu_i - \mu_C > 0$ for i = 1, 2,…, K – 1 where μ_i represents the mean of the i^{th} sample population and μ_C represents the mean of the control population.

M-6.6.1.3. Let α represent the chosen significance level for the test.

M-6.6.1.4. Verify the assumptions of normality. For each sample population, make sure that $0.5 < m/n_i < 2$. If not, Dunnett's Test should *not* be used.

M-6.6.1.5. Calculate the sample mean, $\bar{x}_i$, and the variance, s_i^2, for each of the K – 1 populations and the control ($i = 1,\ 2, \ldots,\ K-1,\ C$).

M-6.6.1.6. Calculate the pooled standard deviation:

$$s_p = \sqrt{\frac{(m-1)s_C^2 + (n_1-1)s_1^2 + \ldots + (n_{K-1}-1)s_{K-1}^2}{(m-1) + (n_1-1) + \ldots + (n_{K-1}-1)}} \ .$$

For each of the K – 1 sample populations, compute

$$t_i = \frac{\bar{x}_i - \bar{x}_C}{s_p\sqrt{1/n_i + 1/n_C}} \ .$$

M-6.6.1.7. Use Table B-26 of Appendix B to determine the critical value, $t_{1-\alpha, v_E}$, where the degrees of freedom v_E = $(m-1) + (n_1-1) + \ldots + (n_{K-1}-1)$. Compare t_i to $t_{1-\alpha, v_E}$ for each of the K – 1 sample populations.

M-6.6.1.7.1. If t_i > $t_{1-\alpha, v_E}$ for any sample population, then reject H_0 and conclude that the mean of the sample population exceeds the mean of the control population.

M-6.6.1.7.2. Otherwise, conclude that the mean of the sample population does not exceed the mean of the control population.

M-6.6.2. Example of Dunnett’s Test for Simple Random and Systematic Samples. Suppose manganese (Mn) groundwater concentrations at six wells are going to be compared to a background well at Site A using the following test with 95% level of confidence.

M-6.6.2.1. H_0: $\mu_i - \mu_C \le 0$ (no difference between the sample means and the control mean).

M-6.6.2.2. H_A: $\mu_i - \mu_C > 0$ for $i = 1, 2,\ldots, K - 1$ where μ_i represents the mean of the i^{th} sample population and μ_C represents the mean of the control population.

M-6.6.2.3. All Mn concentrations were detected so no proxy concentrations are needed to evaluate the data.

M-6.6.2.4. The assumptions of normality were verified during the ANOVA process. Because the sample population for each well is equal to 8, we only have to calculate m/n_i once. As $m/n_i = 8/8 = 1$ is between 0.5 and 2, it is reasonable to apply Dunnett’s test.

Well	69-2-02	69-2-04	69-2-08	69-2-05	69-2-06B	69-2-06A	Bkgd
Mean of Log Result	–0.832	–2.877	–0.907	–5.018	–2.008	–1.144	–4.317
Variance	0.064	0.041	0.091	0.033	0.143	0.011	0.080

$$s_p = \sqrt{\frac{(m-1)s_c^2 + (n_1-1)s_1^2 + \ldots + (n_{K-1}-1)s_{K-1}^2}{(m-1)+(n_1-1)+\ldots+(n_{K-1}-1)}}$$

$$= \sqrt{\frac{7(0.080+0.064+0.041+0.091+0.033+0.143+0.011)}{7+7+7+7+7+7+7}} = \sqrt{\frac{3.240}{49}} = 0.2571$$

$$t_i = \frac{\bar{x}_i - \bar{x}_C}{s_p\sqrt{1/n_i + 1/n_C}} = \frac{\bar{x}_i - (-4.317)}{0.2571\sqrt{1/8+1/8}} = \frac{\bar{x}_i + 4.317}{0.1286}$$

so for each sample well

Sample Well, *i*	$\boldsymbol{t_i}$
69-2-02	27.11
69-2-04	11.20
69-2-08	26.52
69-2-05	–5.45
69-2-06B	17.96
69-2-06A	24.68

M-6.6.2.5. The degrees of freedom are $(8-1)+(8-1)+\ldots+(8-1)=49$. So, using Table B-26 of Appendix B with 49 degrees of freedom, the critical value $t_{0.95,49}$ = 2.32.

M-6.6.2.5. For all wells except Well 69-2-05, $t_i > t_{0.95,\,49}$. We then reject H_0 and conclude that the means of the sample well populations exceed the mean of the control well population, except for Well 69-2-05.

Table M-16.
Data for Example M-6.6.2

Well Location	Result (mg/L)	Log Result	Well Location	Result (mg/L)	Log Result
69-2-02	0.432	–0.839	69-2-06A	0.294	–1.224
69-2-02	0.44	–0.821	69-2-06A	0.301	–1.201
69-2-02	0.513	–0.667	69-2-06A	0.379	–0.970
69-2-02	0.704	–0.351	69-2-06A	0.352	–1.044
69-2-02	0.327	–1.118	69-2-06A	0.346	–1.061
69-2-02	0.316	–1.152	69-2-06B	0.13	–2.040
69-2-02	0.454	–0.790	69-2-06B	0.184	–1.693
69-2-02	0.401	–0.914	69-2-06B	0.209	–1.565
69-2-04	0.0504	–2.988	69-2-06B	0.2	–1.609
69-2-04	0.0502	–2.992	69-2-06B	0.0739	–2.605
69-2-04	0.054	–2.919	69-2-06B	0.0876	–2.435
69-2-04	0.0523	–2.951	69-2-06B	0.126	–2.071
69-2-04	0.0923	–2.383	69-2-06B	0.129	–2.048
69-2-04	0.0556	–2.890	bkgd	0.0137	–4.290
69-2-04	0.0534	–2.930	bkgd	0.019	–3.963
69-2-04	0.0517	–2.962	bkgd	0.0163	–4.117
69-2-05	0.00684	–4.985	bkgd	0.0195	–3.937
69-2-05	0.00639	–5.053	bkgd	0.0112	–4.492
69-2-05	0.00631	–5.066	bkgd	0.0112	–4.492
69-2-05	0.00813	–4.812	bkgd	0.0102	–4.585
69-2-05	0.00747	–4.897	bkgd	0.00946	–4.661
69-2-05	0.00679	–4.992	69-2-08	0.563	–0.574
69-2-05	0.00731	–4.919	69-2-08	0.512	–0.669
69-2-05	0.00444	–5.417	69-2-08	0.475	–0.744
69-2-06A	0.3	–1.204	69-2-08	0.546	–0.605
69-2-06A	0.286	–1.252	69-2-08	0.276	–1.287
69-2-06A	0.303	–1.194	69-2-08	0.383	–0.960
			69-2-08	0.33	–1.109
			69-2-08	0.27	–1.309

M-6.7. Scheffé’s Test. Scheffé’s test is designed to allow the comparison of any set of contrasts while controlling the experiment-wise Type I error rate (the probability of declaring any contrast different from 0 when it is not) to be no more than α (Montgomery, 1997). When the experimenter is only interested in comparing pairs of treatment means, Scheffé's test is not the most sensitive. Directions for Scheffé's Test and an example are presented in Paragraphs M-6.7.1 and M-6.7.2, respectively.

M-6.7.1. Directions for Scheffé's Test. Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let

$$N = \sum_{i=1}^{K} n_i$$

be the overall sample size. Let the values from each population be represented by $x_{i,j}$ where *i* = 1, 2,…, *K* for the *K* groups and $j = 1, 2, ..., n_i$ for the observations in the *i*th group. Let $(1-\alpha)100\%$ be the confidence level for the test.

M-6.7.1.1. Verify the assumptions of normality. Let

$$\theta = \sum a_i \mu_i$$

represent one of *m* linear combinations of the means u_i being tested for $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$.

M-6.7.1.2. Reject H_0 if $|\theta| = \left| \sum a_i \bar{x}_i \right|$ exceeds the critical value

$$S_\alpha = \sqrt{MSE \sum_{i=1}^{K} \left(a_i^2 / n_i \right)} \sqrt{(K-1) F_{1-\alpha, K-1, N-K}}$$

where n_i is the number of observations in the *i*th group of

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j}$$

and

$$F_{1-\alpha, K-1, N-K}$$

is the $(1-\alpha)100\%$ percentile for the *F* distribution with *K* – 1 numerator degrees of freedom and *N* – *K* denominator degrees of freedom (see Table B-7 in Appendix B).

M-6.7.2. Example of Scheffé's Test. Suppose manganese concentrations in groundwater are going to be compared in six different sampling wells and a background well using Scheffé's test with a 95% level of confidence.

M-6.7.2.1. Table M-16 presents the data. All manganese concentrations were detected, so no proxy concentrations are needed to evaluate the data. The assumptions of normality were verified during the ANOVA process.

M-6.7.2.2. Suppose two contrasts are of interest: comparing the background well to all of the other wells combined and comparing well 69-2-06A to well 69-2-06B. These two contrasts can be written:

$$\theta_1 = 6\mu_{bkgd} - \mu_{69-2-02} - \mu_{69-2-04} - \mu_{69-2-05} - \mu_{69-2-06A} - \mu_{69-2-06B} - \mu_{69-2-08}$$

$$\theta_2 = \mu_{69-2-06A} - \mu_{69-2-06B} \ .$$

The contrast estimates are:

$$\hat{\theta}_1 = 6\bar{x}_{bkgd} - \bar{x}_{69-2-02} - \bar{x}_{69-2-04} - \bar{x}_{69-2-05} - \bar{x}_{69-2-06A} - \bar{x}_{69-2-06B} - \bar{x}_{69-2-08}$$

$$= 6(-4.317) - (-0.832) - (-2.877) - (-5.018) - (-1.144) - (-2.008) - (-0.907)$$
$$= -13.1177$$

$$\hat{\theta}_2 = \bar{x}_{69-2-06A} - \bar{x}_{69-2-06B} = -1.144 - (-2.008) = 0.8646.$$

The critical values are:

$$S_{\alpha_1} = \sqrt{\mathrm{MSE}\sum_{i=1}^{K}\left(a_i^2 / n_i\right)} \times \sqrt{(K-1)\,F_{1-\alpha,K-1,N-K}}$$
$$= \sqrt{0.066 \times \left(\frac{36}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right)} \times \sqrt{(7-1)F_{0.95,6,49}}$$
$$= 0.589 \times \sqrt{6 \times 2.29} = 2.1841$$

$$S_{\alpha_2} = \sqrt{\mathrm{MSE}\sum_{i=1}^{K}\left(a_i^2 / n_i\right)} \times \sqrt{(K-1)\,F_{1-\alpha,K-1,N-K}}$$
$$= \sqrt{0.066 \times \left(\frac{1}{8} + \frac{1}{8}\right)} \times \sqrt{(7-1)\,F_{0.95,6,49}}$$
$$= 0.128 \times \sqrt{6 \times 2.29} = 0.4766 \ .$$

M-6.7.2.3. Because the absolute value of each contrast exceeds the relevant critical value, we reject $H_0 : \theta_1 = 0$ and $H_0 : \theta_2 = 0$ with 95% confidence. In other words, the average measurement at the background well is significantly different from the average measurement at the other six wells, and the average measurement at well 69-2-06A differs significantly from the average at well 69-2-06B.

APPENDIX N

Hypothesis Testing—Tests of Dispersion

N-1. Introduction. Many statistical tests make assumptions on the dispersion of data, as measured by variance. This Appendix considers some of the most commonly used statistical tests for equality of variance, a key assumption for the validity of a two-sample *t*-test and analysis of variance (ANOVA). More information on hypothesis tests on the variance can be found in EPA/240/B-026/003, QA/G-9S.

N-2. F-Test for the Equality of Two Variances. An *F*-test may be used to see 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. (In the case of non-normality, Levene's test is recommended, Paragraph N-4.) Directions for implementing an *F*-test are given in Paragraph N-2.1, followed by an example in Paragraph N-2.2.

N-2.1. Directions for an F-Test Comparing Two Variances. 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 *n* data points from population 2. To perform an *F*-test, proceed as follows.

N-2.1.1. Test the null hypothesis of equal variances:

$$H_0 : \sigma_x^2 = \sigma_y^2,\ H_A : \sigma_x^2 \neq \sigma_y^2 \ .$$

N-2.1.2. Verify the assumption of normality using one of the methods described in Appendix F.

N-2.1.3. Calculate the sample variance, s_x^2 (for the X's) and s_y^2 (for the Y's) (Appendix D).

N-2.1.4. Calculate the variance ratios, $f_x = s_x^2 / s_y^2$ and $f_y = s_y^2 / s_x^2$.

N-2.1.5. Let *f* equal the larger of these two values.

N-2.1.5.1. If $f = f_x$, then let $k = m - 1$ and $q = n - 1$.

N-2.1.5.2. If $f = f_y$, then let $k = n - 1$ and $q = m - 1$.

N-2.1.6. Using Table B-7 of Appendix B of the *F*-distribution, we find the critical value,

$$U = F_{1-\alpha/2,k,q}$$

Where *k* denotes the degrees of freedom in the numerator and *q* the degrees of freedom in the denominator for the ratio *f*.

N-2.1.6.1. If $f > U$, conclude that the variances of the two populations are not the same.

N-2.1.6.2. If $f \le U$, there is insufficient evidence to conclude the variances are different.

N-2.2. Example of an F-Test Comparing Two Variances. Consider the case where nickel concentrations in surface soil are compared between Site A and Background. The null and alternative hypotheses are:

$$H_0 : \sigma_x^2 = \sigma_y^2,\ H_A : \sigma_x^2 \ne \sigma_y^2\ .$$

N-2.2.1. Nickel in surface soils at Site A (*X*) was detected at following concentrations (*m* = 6): 2.665, 3.610, 5.470, 7.150, 8.340, 7.960 mg/kg.

N-2.2.2. Nickel in surface background (bkgd) soils (*Y*) was detected at the following concentrations (*n* =10): 5.140, 7.460, 5.990, 3.360, 3.190, 2.870, 5.950, 1.720, 4.770, 5.605 mg/kg.

N-2.2.3. Verify the assumption of normality. For this case, the Shapiro-Wilk test is used.

N-2.2.4. Calculate the sample variance, s_x^2 (for the $X's$) and s_y^2 (for the $Y's$).

Data	**Sample Mean**	**Sample Variance**	**Sample Size**
Site	5.87	5.53	6
Background	4.61	3.12	10

N-2.2.5. Calculate the variance ratios:

$$f_x = s_x^2 / s_y^2 = \frac{5.53}{3.12} = 1.77$$

and

$f_y = s_y^2 / s_x^2 = \dfrac{3.12}{5.53} = 0.56$.

N-2.2.6. Therefore $f = 1.77$.

N-2.2.7. Because $f = f_x$, $k = 6 - 1 = 5$ and $q = 10 - 1 = 9$.

$U = F_{1-\alpha/2,k,q} = f_{0.975,5,9} = 4.484$.

N-2.2.8. Because $f \le U$ $(1.77 \le 4.484)$, there is insufficient evidence to conclude the variances are different.

N-3. Bartlett's Test for the Equality of Two or More Variances. Bartlett's test, which is essentially a generalization of the *F*-test, is a way 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. Directions for Bartlett's test are given in Paragraph N-3.1, followed by an example in Paragraph N-3.1. Like the *F*-test it is sensitive to deviations from normality.

N-3.1. Directions for Bartlett's Test for Two or More Variances. Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let *N* represent the total number of samples, $N = n_1 + n_2 .... + n_k$. Let the values from each population be represented by $x_{i,j}$, where $i = 1, 2, \ldots, K$ for the *K* groups and $j = 1, 2, \ldots, n_i$ for the observations in the *i*th group.

N-3.1.1. $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$ (no difference among the population variances).

N-3.1.2. H_A: at least one variance, σ_i^2, is different from one or more of the other variances.

N-3.1.3. For example, consider two wells, where four samples have been taken from Well 1 and three samples have been taken from Well 2. In this case, $K = 2$, $n_1 = 4$, $n_2 = 3$, and $N = 4 + 3 = 7$.

N-3.1.4. Verify the assumption of normality using one of the methods described in Appendix F. For each of the *K* groups, calculate the sample variances, s_i^2 (see Appendix D).

N-3.1.5. Compute the pooled variance using the *K* groups:

$$s_p{}^2 = \frac{1}{(N-K)}\sum_{i=1}^{K}(n_i - 1)s_i{}^2 \ .$$

N-3.1.6. Compute the test statistic (*TS*):

$$TS = (N-K)\mathrm{Ln}(s_p{}^2) - \sum_{i=1}^{K}(n_i - 1)\mathrm{Ln}(s_i{}^2)$$, where Ln is the natural logarithm.

N-3.1.7. Using a chi-square table (Table B-2 of Appendix B), find the critical value of the chi-squared distribution, $\chi^2_{1-\alpha,\nu}$, with $\nu = K-1$ degrees of freedom and the $(1-\alpha)100\%$ level of confidence. For example, for a level of confidence of 95% (significance level α = 0.05) and ν = 5, $\chi^2_{0.95,5}$ = 11.1.

N-3.1.7.1. If $TS > \chi^2_{1-\alpha,\nu}$, reject H_0 (conclude that the variances are not all equal) at the $(1-\alpha)100\%$ level of confidence.

N-3.1.7.2. If $TS \leq \chi^2_{1-\alpha,\nu}$, there is insufficient evidence to reject H_0.

N-3.2. <u>Example of Bartlett's Test for Two or More Variables</u>. Using chromium concentrations in subsurface site soil, the data are: 2.95, 5.17, 4.80, 4.53, 4.01, 5.91, 3.96, 4.81, 5.27, 5.99, 4.60, 5.51, 4.72, 3.56, 4.22, 3.91, 5.81, 4.48, 5.10, 4.94, 4.76, 4.62, 4.72, 4.73, 3.21, 4.14, 4.85, 4.25, 5.09, 3.68, 5.12, 6.60, 6.19, 3.15, 4.11, 2.80 mg/kg.

N-3.2.1. The chromium concentrations in subsurface background soil are: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, 3.84 mg/kg.

N-3.2.2. Verify the assumption of normality. For this case, the Shapiro-Wilk test is used.

N-3.2.3. Let *N* represent the total number of samples. As the site data has $n_1 = 36$ samples and the background data has $n_2 = 8$ samples, $N = 44$ and $K = 2$.

N-3.2.4. For each of the *K* groups, calculate the sample variances, s_i^2: s_1^2 = 0.806 (site variance) and s_2^2 = 0.526 (background variance).

N-3.2.5. Compute the pooled variance:

$$s_p^2 = \frac{1}{(N-K)}\sum_{i-1}^{K}(n_i - 1)s_i^2$$

$$= \frac{1}{(44-2)}\left[(n_1-1)s_1^2+(n_2-1)s_2^2\right]$$

$$= \frac{1}{42}\left[(36-1)\,0.806\ +\ (8-1)\,0.526\right] = 0.7593\ .$$

N-3.2.6. Compute the test statistic *TS*:

$$TS = (N-K)\,\text{Ln}(s_p^2) - \sum_{i=1}^{K}(n_i-1)\,\text{Ln}(s_i^2)$$

$$= (44-2)\,\text{Ln}(0.7593) - \left[(36-1)\,\text{Ln}(0.806) + (8-1)\,\text{Ln}(0.526)\right] = 0.4802\ \ .$$

N-3.2.7. Using a chi-squared table (Table B-2 of Appendix B), find the critical value, $\chi^2_{1-\alpha,\nu}$. In this case, with a significance level of 5% and 1 degree of freedom, $\chi^2_{0.95,1}$= 3.841. As $TS = 0.4802 \le 3.841$, there is insufficient evidence to conclude the variances are different at the $\alpha = 0.05$ significance level.

N-4. Levene's Test for the Equality of Two or More Variances. Levene's test is a non-parameter 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 because it involves applying an ANOVA to the absolute deviations from the group means. Directions for Levene's test are given in Paragraph N-4.1, followed by an example in Paragraph N-4.2.

N-4.1. Directions for Levene's Test for the Equality of Two or More Variances. Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let *N* represent the total number of samples, $N = n_1 + n_2 .... + n_k$. Let the values from each population be represented by $x_{i,j}$ where $i = 1,2,\ldots,K$ for the *K* groups and $j = 1,2,\ldots,n_i$ for the observations in the *i*th group.

N-4.1.1. $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$ (no difference among the population variances).

N-4.1.2. H_A: at least one variance, σ_i^2, is different from one or more of the other variances.

N-4.1.3. For example, consider two wells where four samples have been taken from well 1 and three samples have been taken from well 2. In this case, $K = 2$, $n_1 = 4$, $n_2 = 3$, and $N = 4 + 3 = 7$.

N-4.1.4. Verify the assumption of normality using one of the methods described in Appendix F. For each of the *K* groups, calculate the group mean, $\overline{x}_i$:

$$\overline{x}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1,\,j} \ , \quad \overline{x}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2,\,j} \ , \ \ldots. \ , \quad \overline{x}_K = \frac{1}{n_K}\sum_{j=1}^{n_K} x_{K,\,j} \ .$$

N-4.1.5. Compute the absolute residuals

$$z_{i,\,j} = \left| x_{i,\,j} - \overline{x}_i \right|$$

where $x_{i,\,j}$ represents the *j*th value of the *i*th group. For each of the *K* groups, calculate the means, $\overline{z}_i$,of these residuals:

$$\overline{z}_1 = \frac{1}{n_1}\sum_{j=1}^{n_2} z_{1,\,j} \ , \quad \overline{z}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} z_{2,\,j} \ , \ldots., \quad \overline{z}_K = \frac{1}{n_K}\sum_{j=1}^{n_K} z_{K,\,j} \ .$$

N-4.1.6. Calculate the overall mean residual:

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i} z_{i,\,j} = \frac{1}{n}\sum_{i=1}^{K} n_i \overline{z}_i \ .$$

N-4.1.7. Compute the following sums of squares for the absolute residuals:

$$SS_{TOTAL} = \sum_{i=l}^{K}\sum_{j=1}^{n_i} z_{i,\,j}^2 - n\,\overline{z}^2$$

$$SS_{GROUPS} = \sum_{i=l}^{K} \frac{\overline{z}_i^{\,2}}{n_i} - n\,\overline{z}^2$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{GROUPS} \ .$$

N-4.1.8. Compute

$$f = \frac{SS_{GROUPS}\ /(K-1)}{SS_{ERROR}\ /(N-K)} \ .$$

N-4.1.9. Using Table B-7 of Appendix B, find $F_{1-\alpha,\,k-1,\,N-K}$, the critical value of the *F*-distribution with $(K - 1)$ numerator degrees of freedom, $(N-K)$ denominator degrees of freedom, and the desired level of significance, α. For example, if $\alpha = 0.05$, the numerator degrees of freedom are 5, and the denominator degrees of freedom are 18, then using Table B-7, we find that $F_{0.95,5,18} = 2.77$.

N-4.1.10. If $f > F$, reject the assumption of equal variances.

N-4.2. Example of Levene's Test for the Equality of Two or More Variables. Consider the case where nickel concentrations in surface soil are compared between Site A and background (bkgd) using the test:

$$H_0 : \sigma_x^2 = \sigma_y^2, \;\; H_A : \sigma_x^2 \neq \sigma_y^2 \;.$$

N-4.2.1. Suppose data for nickel in surface site soil are: 2.665, 3.610, 5.470, 7.150, 8.340, 7.960 mg/kg. And suppose data for nickel in surface background are: 5.140, 7.460, 5.990, 3.360, 3.190, 2.870, 5.950, 1.720, 4.770, 5.605 mg/kg.

N-4.2.2. Verify the assumption of normality. For this case, the Shapiro-Wilk test is used.

Site mean = $\bar{x}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1,j} = 5.87$.

Background mean = $\bar{x}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2,j} = 4.61$.

Site A	$z_{i,j} = \left\| x_{i,j} - \bar{x}_i \right\|$	**Background**	$z_{i,j} = \left\| x_{i,j} - \bar{x}_i \right\|$
2.67	3.20	5.14	0.534
3.61	2.26	7.46	2.854
5.47	0.40	5.99	1.384
7.15	1.28	3.36	1.246
8.34	2.47	3.19	1.416
7.96	2.09	2.87	1.736
		5.95	1.344
		1.72	2.886
		4.77	0.164
		5.61	0.999

Mean of the site residuals = $\bar{z}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} z_{1,j} = 1.95$.

Mean of the background residuals = $\bar{z}_2 = \dfrac{1}{n_2}\sum_{j=1}^{n_2} z_{2,\,j} = 1.46$.

Overall residual mean = $\bar{z} = \dfrac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i} z_{i,\,j} = \dfrac{1}{n}\sum_{i=1}^{K} n_i\,\bar{z}_i = 1.64$.

$$SS_{TOTAL} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} z_{i,\,j}^2 - n\,\bar{z} = 12.60\,.$$

$$SS_{GROUPS} = \sum_{i=l}^{K}\frac{\bar{z}_i^{\;2}}{n_i} - n\,\bar{z} = 0.92\,.$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{GROUPS} = 11.68\,.$$

$$f = \frac{SS_{GROUPS}\,/(K-1)}{SS_{ERROR}\,/(N-K)} = \frac{0.9167\,/(2-1)}{11.68\,/(16-2)} = 1.098\,.$$

Numerator degrees of freedom: (*K* – 1) = (2 – 1) = 1.

Denominator degrees of freedom: (*N* – *K*) = (16 – 2) = 14.

N-4.2.3. Because $\alpha = 0.05$, the critical value $F_{0.95,1,14} = 4.611$. Comparing the calculated value (f) and the critical value, $F_{0.95,1,14}$, we see that $f \le F_{0.95,1,14}$, so do not reject H_0. Therefore, we can conclude that the variance for the surface soil site concentration of nickel is equal to the variance of the surface soil background concentrations of nickel.

N-5. Maximum *F*-Ratio Test for Equality of Two or More Variances. The maximum *F*-ratio tests whether three or more population variances from normal distributions are equal (Mason et al., 1989). The test also assumes that the sample sizes for the populations are equal. As this test is sensitive to departures from normality, it is recommended that normality tests be done before using it. Directions are given in Paragraph N-5.1, followed by an example in Paragraph N-5.2.

N-5.1. Directions for the Maximum F-Ratio Test for Equality of Two or More Variances. Let *K* represent the total number of populations to be compared. Let $n_1, n_2, \ldots, n_K$ represent the sample sizes of each of the *K* sample populations. Let *N* represent the total number of samples, $N = n_1 + n_2 .... + n_k$. Let the values from each population be represented by $x_{i,\,j}$ where $i = 1, 2, \ldots, K$ for the *K* groups and $j = 1, 2, \ldots, n_i$ for the observations in the *i*th group.

N-5.1.1. $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$ (no difference among the population variances).

N-5.1.2. H_A: at least one variance, σ_i^2, is different from one or more of the other variances.

N-5.1.3. Verify the assumption of normality using one of the methods described in Appendix F.

N-5.1.4. Calculate the sample standard deviation for each of the *K* data sets. Denote these standard deviations by s_i and the corresponding sample size by n_i, where $i = 1, 2, \ldots, K$. Identify the largest value of s_i, max(s_i), and the smallest value of s_i, min(s_i).

N-5.1.5. Calculate the ratio $f_{\max} = \left(\max(s_i)/\min(s_i)\right)^2$.

N-5.1.6. If $n_1 = n_2 = \ldots = n_k = \ldots = n$, use the critical values in Table B-27 of Appendix B, $F_{k,v,\alpha}$, where $k = K$ and $v = n - 1$ for the desired level of significance α, to determine whether to reject the hypothesis of equal standard deviations. If the n_i is unequal but not too different, use the "harmonic mean of the n_i", n':

$$v = n' - 1, \text{ where } n' = K \Big/ \sum_{i=1}^{K} \left(1/n_i\right).$$

N-5.1.7. If $f_{\max} > F_{K,v,\alpha}$, then conclude there is evidence that the variances are not equal.

N-5.2. Example of the Maximum F-Ratio Test for Equality of Three or More Variances. Manganese concentrations in groundwater are compared between seven wells from Site A using the test:

N-5.2.1 $H_0 : \sigma_i^2 = \sigma_j^2$ for all *i* and *j.*

N-5.2.2. $H_A : \sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$.

N-5.2.3. The data (Table N-1) were tested for equal variances using Bartlett's test (Paragraph N-3). The data were also tested for normality using the Shapiro-Wilk test. Because the data were not normal, they were transformed so that residuals would follow a normal distribution.

Well	69-2-02	69-2-04	69-2-05	69-2-06A	69-2-06B	69-2-07	69-2-08
s_i	0.254	0.203	0.182	0.103	0.378	0.283	0.301.

$$f_{\max} = \left(\max(s_i)/\min(s_i)\right)^2 = (0.301/0.103)^2 = 8.54 .$$

Table N-1.
Data for Example N-5.2

	Result	**Log**		**Result**	**Log**
Well Location	**(mg/L)**	**Result**	**Well Location**	**(mg/L)**	**Result**
69-2-02	0.432	–0.839	69-2-06A	0.294	–1.224
69-2-02	0.44	–0.821	69-2-06A	0.301	–1.201
69-2-02	0.513	–0.667	69-2-06A	0.379	–0.970
69-2-02	0.704	–0.351	69-2-06A	0.352	–1.044
69-2-02	0.327	–1.118	69-2-06A	0.346	–1.061
69-2-02	0.316	–1.152	69-2-06B	0.13	–2.040
69-2-02	0.454	–0.790	69-2-06B	0.184	–1.693
69-2-02	0.401	–0.914	69-2-06B	0.209	–1.565
69-2-04	0.0504	–2.988	69-2-06B	0.2	–1.609
69-2-04	0.0502	–2.992	69-2-06B	0.0739	–2.605
69-2-04	0.054	–2.919	69-2-06B	0.0876	–2.435
69-2-04	0.0523	–2.951	69-2-06B	0.126	–2.071
69-2-04	0.0923	–2.383	69-2-06B	0.129	–2.048
69-2-04	0.0556	–2.890	69-2-07	0.0137	–4.290
69-2-04	0.0534	–2.930	69-2-07	0.019	–3.963
69-2-04	0.0517	–2.962	69-2-07	0.0163	–4.117
69-2-05	0.00684	–4.985	69-2-07	0.0195	–3.937
69-2-05	0.00639	–5.053	69-2-07	0.0112	–4.492
69-2-05	0.00631	–5.066	69-2-07	0.0112	–4.492
69-2-05	0.00813	–4.812	69-2-07	0.0102	–4.585
69-2-05	0.00747	–4.897	69-2-07	0.00946	–4.661
69-2-05	0.00679	–4.992	69-2-08	0.563	–0.574
69-2-05	0.00731	–4.919	69-2-08	0.512	–0.669
69-2-05	0.00444	–5.417	69-2-08	0.475	–0.744
69-2-06A	0.3	–1.204	69-2-08	0.546	–0.605
69-2-06A	0.286	–1.252	69-2-08	0.276	–1.287
69-2-06A	0.303	–1.194	69-2-08	0.383	–0.960
			69-2-08	0.33	–1.109
			69-2-08	0.27	–1.309

N-5.2.4. Because $n_1 = n_2 = \ldots = n_k = \ldots = n$, use the critical values in Table B-27 of Appendix B with $v = n - 1 = 8 - 1 = 7$. So, $F_{K,v,\alpha} = F_{7,7,0.05} = 11.80$.

N-5.2.5. Compare the calculated value (8.54) to the critical value (11.80); because the calculated value $f_{\max}$ is not greater than the critical value, H_0 cannot be rejected (i.e., there is evidence that the variances are equal).

APPENDIX O

Measures of Correlation

O-1. Introduction. A correlation coefficient provides a measure of the degree of association between two variables or measurements. For example, the degree of association between pH and the concentration of a dissolved metal in groundwater may be of interest. The primary objective of calculating a correlation coefficient is to determine whether one variable increases or decreases as the second variable increases, or whether the two variables vary independently of one another.

O-1.1. In environmental applications, a correlation coefficient may be used to determine the strength of an association. For example, numerous groundwater sites contaminated with chlorinated solvents also have high dissolved iron concentrations. Is it possible to determine whether the high iron locations are the same as where chlorinated solvent levels are also high? A correlation coefficient for the relationship provides a quantitative measure of the degree of association of these measured parameters.

O-1.2. A high correlation coefficient does not prove cause and effect. When the correlation between two variables is high, the relationship is strong; but one cannot conclude that one variable causes the other variable to increase or decrease without further evidence. Measuring and identifying correlation is often critical for environmental data, which are frequently correlated over time or space, or both.

O-1.3. Classical statistical methods typically assume data are not correlated. If correlations are not identified before data are statistically evaluated, then statistical methods can provide misleading results. There are also statistics that depend upon correlation in the data, such as geostatistics (Appendix R), and there are methods available for “detrending” or “uncorrelating” data under certain circumstances. These cases are beyond the scope of this discussion, and may be best addressed by a statistician.

O-1.4. Several different correlation coefficients for measuring the degree of association between two variables will be discussed. The correlation coefficients share common properties. Each is a dimensionless quantity with values ranging from –1 to 1. A positive correlation coefficient for two variables indicates that one variable tends to increase as the other variable increases. A negative correlation indicates that one variable tends to decrease as the other variable increases. The highest possible degree of correlation occurs when the absolute value of the correlation coefficient equals one. When two variables are truly independent, the behavior of one variable cannot be predicted from the other variable, and the correlation coefficient is zero. The references EPA/240/B-026/003, QA/G-9S and Conover (1980) contain additional details about measures of correlation.

O-2. Correlation Coefficients as Hypothesis Testing.

O-2.1. Introduction. Calculated values of a correlation coefficient for a set of actual measurements are rarely identically equal to zero when a true correlation is absent (when the true correlation coefficient $\gamma = 0$). Therefore, a hypothesis test is done to determine the presence or absence of a significant correlation. Hypothesis tests are discussed in additional detail in Appendices L, M, and N.

O-2.1.1. The significance of the correlation is often evaluated using a hypothesis test in the form:

H_0: $\gamma = 0$, H_A: $\gamma \neq 0$.

O-2.1.1.1. The correlation coefficient for a set of measured results (x_i, y_i) is initially calculated. The calculated (sample) correlation coefficient, $\hat{\gamma}$, is viewed as an approximation of the population correlation coefficient, γ, for the *X* and *Y* variables.

O-2.1.1.2. The probability, *p*, of obtaining the calculated value when *X* and *Y* are not correlated (when the true correlation coefficient $\gamma = 0$) is then determined. The probability is typically calculated by statistical software.

O-2.1.1.3. If *p* is sufficiently small (e.g., $p \leq \alpha = 0.05$ or 0.01), then a correlation exists. More accurately, the null hypothesis that the true correlation coefficient is zero is rejected (with a level of confidence of at least $1 - \alpha$).

O-2.1.1.4. When statistical software is unavailable, the largest possible absolute value of a correlation coefficient that can occur when *X* and *Y* are not correlated is obtained from a table. The tabular value for the $1 - \alpha$ level of confidence is subsequently compared to the calculated value. If the calculated value is larger than the value obtained from the table, the null hypothesis is rejected, and the correlation coefficient is not equal to zero.

O-2.1.2. Directions and an example for using a correlation coefficient statistical test are in Paragraphs O-2.2 and O-2.3, respectively.

O-2.1.3. Typically, a correlation coefficient is viewed to be significantly different from zero if the *p* value is less than a specified significance level, usually taken to be between 0.1 and 0.01. The *p* value is discussed in more detail in Appendices L, M, and N. Various values for the absolute value of the correlation coefficient, $|\gamma|$, qualitatively describe the degree of association below:

Absolute value of correlation coefficient	**Degree of relationship**
$\lvert\gamma\rvert < 0.50$	Extremely Weak
$0.50 < \lvert\gamma\rvert < 0.75$	Weak
$0.75 < \lvert\gamma\rvert < 0.90$	Moderate
$0.90 < \lvert\gamma\rvert < 0.95$	Moderately Strong
$0.95 < \lvert\gamma\rvert < 1.00$	Strong

O-2.1.4. Four different sample correlation coefficients are discussed below.

O-2.1.4.1. Pearson's *r*.

O-2.1.4.2. Spearman's rho (ρ).

O-2.1.4.3. Serial correlation coefficient.

O-2.1.4.4. Kendall's tau (τ).

O-2.1.5. Pearson's *r* measures the degree of correlation between two variables for linear relationships. Kendall's τ and Spearman's ρ measure the degree of any monotonic relationship between two variables. Two variables, *X* and *Y*, are monotonically correlated if, overall, *Y* consistently increases or decreases as *X* increases. Note that *X* and *Y* will not be monotonically correlated if, as *X* increases, *Y* increases then decreases (or decreases then increases).

O-2.2. Directions for a Correlation Coefficient Statistical Test. Calculate the test statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}.$$

O-2.2.1. Use Table B-23 of Appendix B to find the critical value $t_{1-\alpha/2,\nu}$, which is $(1 - \alpha/2)100^{\text{th}}$ percentile of the Student's *t* distribution with degrees of freedom $\nu = n - 2$.

O-2.2.1.1. Conclude that the correlation is significantly different from zero if

$$\lvert t \rvert > t_{1-\alpha/2,\nu}.$$

O-2.2.1.2. Otherwise, state that there is insufficient evidence to conclude that the correlation coefficient is different from zero.

O-2.2.2. A one-tailed test can be performed in a similar manner by replacing $\alpha/2$ by α. For example, to test whether a correlation exceeds zero, compare *t* with $t_{1-\alpha,n-2}$. If $t > t_{1-\alpha,n-2}$ conclude that the correlation is larger than zero. Otherwise conclude that the true correlation may be less than or equal to zero.

O-2.3. Example of a Test for a Correlation Coefficient. Consider the following data set for chromium and lead in subsurface soil background (in mg/kg).

Sample	**Chromium (*X*)**	**Lead (*Y*)**
EPC-BG01	4.60	3.50
EPC-BG02	5.29	4.16
EPC-BG03	4.26	4.19
EPC-BG04	5.28	3.91
EPC-BG05	4.53	3.66
EPC-BG06	5.74	4.31
EPC-BG07	5.86	4.19
EPC-BG08	3.84	3.35

O-2.3.1. The objective is to test if the correlation coefficient is different from zero, based on 90% level of confidence.

O-2.3.2. For 90% confidence, $\alpha = 0.10$.

O-2.3.3. The correlation coefficient was calculated in Paragraph O-2.4.2 and equals $r = 0.72$.

O-2.3.4. The test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.7229}{\sqrt{\frac{1-(0.7229)^2}{8-2}}} = 2.563$$

with $\nu = 8 - 2 = 6$.

O-2.3.5. The critical value is $t_{1-\alpha/2,n-2} = t_{0.95,6} = 1.943$.

O-2.3.6. Comparing the test statistic to the critical value, *t* = 2.563 > 1.943. With at least 90% confidence, the correlation coefficient is significantly different from zero.

However, given the magnitude of *r*, the linear association between chromium and lead could be qualitatively described as "weak."

O-2.4. Pearson's *r***.** The Pearson's *r* is a parametric measure of correlation for linear relationship between two variables. A linear association implies that, as one variable increases, so does the other in a uniform manner (i.e., linearly), or as one variable decreases the other increases linearly. A value of +1 implies a perfect positive linear correlation, i.e., that all the data pairs (x_i, y_i) lie on a straight line with a positive slope. A value of –1 implies perfect negative linear correlation. Directions and an example for Pearson's correlation coefficient are presented in Paragraphs O-2.4.1 and O-2.4.2.

O-2.4.1. Directions for Pearson's Correlation Coefficient. Let $x_1, x_2, ..., x_n$ represent one variable (X) of the n data points and let $y_1, y_2, ..., y_n$ represent the corresponding values of a second variable (Y). The Pearson correlation coefficient, r, for the sample of (x_i, y_i) pairs is computed by:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\sum_{i=1}^{n} x_i\, y_i - n\,\bar{x}\,\bar{y}}{(n-1)\, s_x\, s_y} .$$

O-2.4.2. Example of Pearson's Correlation Coefficient. Consider the following data set for n = 8 chromium and lead in subsurface soil background (in mg/kg):

Sample	**Chromium(*X*)**	**Lead(*Y*)**
EPC-BG01	4.60	3.50
EPC-BG02	5.29	4.16
EPC-BG03	4.26	4.19
EPC-BG04	5.28	3.91
EPC-BG05	4.53	3.66
EPC-BG06	5.74	4.31
EPC-BG07	5.86	4.19
EPC-BG08	3.84	3.35

O-2.4.2.1. For chromium,

$$\sum_{i=1}^{8} x_i = 4.60 + 5.29 + 4.26 + 5.28 + 4.53 + 5.74 + 5.86 + 3.84 = 39.4$$

$$\sum_{i=1}^{8} x_i^2 = 4.60^2 + 5.29^2 + 4.26^2 + 5.28^2 + 4.53^2 + 5.74^2 + 5.86^2 + 3.84^2 = 197.7 \ .$$

So, $\bar{x} = 39.4/8 = 4.925$ and

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{197.7-(8\times 4.925^2)}{7}} = 0.7226 \ .$$

O-2.4.2.2. For lead,

$$\sum_{i=1}^{8} y_i = 3.50+4.16+4.19+3.91+3.66+4.31+4.19+3.35 = 31.27$$

$$\sum_{i=1}^{8} y_i^2 = 3.50^2+4.16^2+4.19^2+3.91^2+3.66^2+4.31^2+4.19^2+3.35^2 = 123.2\,.$$

So, $\bar{y} = 31.27/8 = 3.909$ and

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1}} = \sqrt{\frac{123.2-(8\times 3.909^2)}{7}} = 0.3632\,.$$

O-2.4.2.3. The “cross term” dependent upon the product of chromium and lead is:

$$\sum_{i}^{n} x_i y_i = (4.60\times 3.50)+(5.29\times 4.16)+(4.26\times 4.19)+(5.28\times 3.91)+$$
$$(4.53\times 3.66)+(5.74\times 4.31)+(5.86\times 4.19)+(3.84\times 3.35) = 155.3.$$

So,

$$r = \frac{155.3-(8\times 4.925\times 3.909)}{7\times 0.7226\times 0.3729} = 0.72\,.$$

Paragraphs O-2.4.3 and O-2.4.4 will demonstrate how to test whether the sample correlation coefficient indicates that the population correlation coefficient differs from zero.

O-2.4.3. Discussion. Although two independent variables will produce a correlation coefficient of zero, it should be noted that a calculated correlation coefficient that is equal to or near zero does not demonstrate the absence of a significant relationship between the two variables. For example, because Pearsons’ *r* does not detect non-linear relationships, a strong non-linear relationship could result in a value of *r* equal to zero.

O-2.4.3.1. The data from the previous example are illustrated in Figure O-1. Correlation coefficients should be used with scatter plots to determine whether a low value of Pearson's *r* is due to a non-linear relationship or a lack of association.

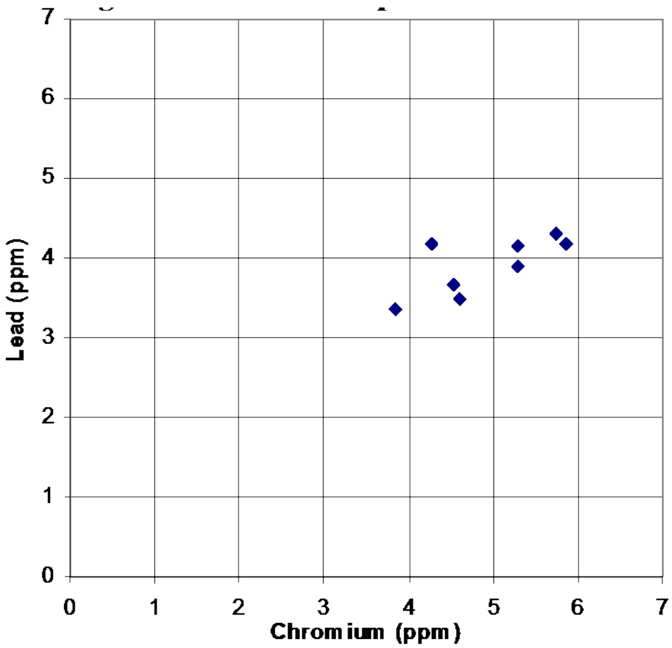


Figure O-1. Scatter Plot for Chromium and Lead.

O-2.4.3.2. Pearson's *r* can be sensitive to the presence of one or two extreme values, especially when sample sizes are small. Such values may result in a high correlation, suggesting a strong linear trend, when only a moderate or weak trend is present. This may happen, for instance, if a single (*x*, *y*) pair has very high values for both measurements while the remaining data values are uncorrelated. For example, Figure O-2 plots an example where a very large outlier exists. Including the outlier leads to a sample correlation coefficient of 0.96. Without this value, the sample correlation coefficient falls to –0.10. Extreme values may also lead to low sample correlation coefficients, thus tending to mask a strong linear trend. This may happen if all the (*x*, *y*) pairs except one (or two) tend to cluster tightly about a straight line, and the exceptional point has a very large *X* value paired with a moderate or small *Y* value (or vice versa). Because of the influences of extreme values, it is wise to use a scatter plot in conjunction with a Pearson correlation coefficient.

O-2.4.3.3. An important property of Pearson's *r* is that it is unaffected by changes in location of the data (adding or subtracting a constant from all of the *X* or *Y* measurements) and by changes in scale of the data (multiplying the *X* or *Y* values by a positive constant). Linear transformations on the data pairs do not affect the correlation coefficient of the measurements. For example, if one variable in the pair was temperature in degrees Celsius, then the correlation would not change if Celsius is converted to Fahrenheit.

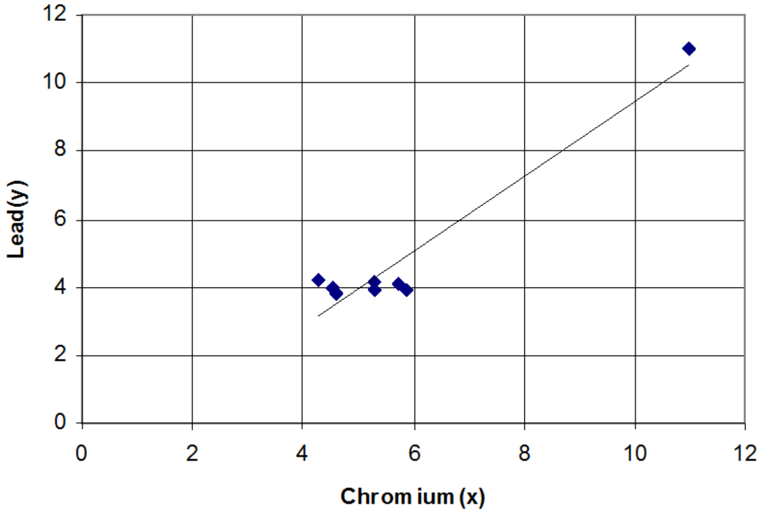


Figure O-2. Scatter Plot with Outlier.

O-2.4.3.4. However, Pearson's *r* is not invariant to non-linear transformations. If non-linear transformations of the measurements are made, then the Pearson correlation coefficient between the transformed values will differ from the Pearson correlation coefficient of the original measurements. For example, if *X* and *Y* represent PCB and dioxin concentrations in soil, respectively, and $U = \text{Log}(X)$ and $V = \text{Log}(Y)$, then the Pearson correlation coefficients between *X* and *Y* and between *U* and *V* will be different because the logarithmic transformation is a nonlinear transformation.

O-2.4.3.5. It should be further noted that statistical tests that use *r* to estimate the population correlation coefficient rely on the assumption that the true relationship between the variables *X* and *Y* follows a bivariate normal distribution. If either variable *X* or *Y* is not normal, then together *X* and *Y* are not likely to follow a bivariate normal distribution. For more details see Snedecor and Cochran (1982).

O-2.5. Spearman's rho. Spearman's rank correlation coefficient measures monotonic correlation for ordinal data (data that can be ranked) and is nonparametric (i.e., can be used when the data are not normally distributed).

O-2.5.1. Introduction. Data may be either linearly or non-linearly correlated. When one variable tends to consistently increase or decrease as another variable increases, the two variables possess a monotonic correlation. Unlike Pearson's *r*, Spearman's rho, ρ, may be used to measure the strength of both linear and nonlinear relationships.

O-2.5.1.1. It is calculated by first replacing each value *x*, by its rank R(*x*) (1 for the smallest *x* value, 2 for the second smallest, etc.) and each value *y* by its rank R(*y*). These pairs of ranks are then treated as the (*x*, *y*) data and Spearman's rank correlation is calculated using the same formula as for Pearson's correlation.

O-2.5.1.2. Directions and an example for calculating a Spearman's rank correlation coefficient are contained in the Paragraphs O-2.5.2 and O-2.5.3.

O-2.5.1.3. Because meaningful (monotonically increasing) transformations of the data will not alter the ranks of the respective variables (the ranks for Log(*x*) will be the same as the ranks for *x*), Spearman's correlation will not be altered by non-linear increasing transformations of *x* and *y*. For instance, the Spearman correlation between PCB and dioxin concentrations (*x* and *y*) in soil will be the same as the correlation between their logarithms, Log(*x*) and Log(*y*). Because Spearman's ρ is a nonparametric measure of correlation, it is invariant for monotonic increasing transformations and is less sensitive to extreme values than Pearson's correlation. However, Pearson's *r* has higher statistical power than Spearman's ρ.

O-2.5.2. Directions for the Spearman's Rank Correlation Coefficient. Let

$$R(x_1), R(x_2), \ldots, R(x_n)$$

represent a set of ranks of the *n* data points for the variable *X* and let

$$R(y_1), R(y_2), \ldots, R(y_n)$$

represent a set of ranks of a second variable *Y* of the *n* data points. The Spearman sample correlation coefficient, ρ, for *X* and *Y* is computed by:

$$\rho = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{R(x_i)-\overline{R}(x)}{s_{R(x)}}\right)\left(\frac{R(y_i)-\overline{R}(y)}{s_{R(y)}}\right) = \frac{\sum_{i=1}^{n} R(x_i)R(y_i) - n\,\overline{R}(x)\overline{R}(y)}{(n-1)\,s_{R(x)}\,s_{R(y)}} .$$

O-2.5.3. Example of Spearman's Correlation Coefficient. Consider the following data set for chromium and lead in subsurface soil background (in mg/kg):

Sample	**Chromium(*X*)**	**Lead(*Y*)**
EPC-BG01	4.60	3.50
EPC-BG02	5.29	4.16
EPC-BG03	4.26	4.19
EPC-BG04	5.28	3.91
EPC-BG05	4.53	3.66
EPC-BG06	5.74	4.31
EPC-BG07	5.86	4.19
EPC-BG08	3.84	3.35

O-2.5.3.1. First the data must be ranked:

Sample	**Chromium**	**Rank (*X*)**	**Lead**	**Rank (*Y*)**
EPC-BG01	4.60	4	3.50	2
EPC-BG02	5.29	6	4.16	5
EPC-BG03	4.26	2	4.19	6.5
EPC-BG04	5.28	5	3.91	4
EPC-BG05	4.53	3	3.66	3
EPC-BG06	5.74	7	4.31	8
EPC-BG07	5.86	8	4.19	6.5
EPC-BG08	3.84	1	3.35	1

O-2.5.3.2. Notice that two of the lead values are equal, so their rank is assigned to be the average of ranks 6 and 7.

O-2.5.3.3. For chromium, $\overline{R}(x) = 4.5$, and $s_{R(x)} = 2.45$.

O-2.5.3.4. For lead, $\overline{R}(y) = 4.5$, and $s_{R(y)} = 2.43$.

O-2.5.3.5. The sum of the cross-products for chromium and lead ranks is:

$$\sum_{i=1}^{8} R(x_i)\,R(y_i) = (1\times1)+(2\times6.5)+(3\times3)+(4\times2)+(5\times4)+(6\times5)+(7\times8)+(8\times6.5)$$
$$=189 \ .$$

O-2.5.3.6. The correlation coefficient is

$$\rho = \frac{189-(8\times4.5\times4.5)}{7\times2.45\times2.43} = 0.647 \ .$$

O-2.6. Serial Correlation Coefficient. The serial correlation coefficient is a measure of the extent to which successive observations (either in time or space) are related. The primary difference between the serial correlation coefficient and other measures of correlation is the manner in which the correlation coefficient is used and the manner in which one of the variables is scaled. For example, the serial correlation coefficient is frequently used to determine the behavior of some variable of interest *X* with respect to time (*t*). Frequently, the variable *X* is measured at equally spaced time intervals, so that the data points are of the form (x_1, t_1), (x_2, t_2),…., (x_n, t_n). The serial correlation coefficient may be a parametric or non-parametric measure of correlation, depending upon how it is calculated. For example, if variable *X* is being evaluated with respect to time *t*, Spearman's ρ is essentially being calculated if the values of *X* are replaced with the corresponding ranks. Directions and examples for calculating a serial correlation coefficient are presented in the following two paragraphs.

O-2.6.1. Directions to Calculate the Serial Correlation Coefficient.

O-2.6.1.1. For a sequence of data points taken serially in time, or "one-by-one in a row," the serial correlation coefficient can be calculated by replacing the sequencing variable by the numbers 1 through *n* and calculating Pearson's correlation coefficient with *x* being the actual data values, and *y* being the numbers 1 through *n*. For example, for a sequence of samples collected every 10 feet along a straight transit line at a waste site, the distances on the transit line of the data points are replaced by the numbers 1 through *n*, for samples taken at 10-foot intervals (first 10-foot sample point = 1, the 20-foot sample point = 2, the 30-foot sample point = 3, etc.).

O-2.6.1.2. To calculate the serial correlation coefficient, let x_1, x_2,..., x_n represent the data values collected in sequence over equally spaced periods. Label the periods 1, 2..., *n* to match the data values. Use the directions above to calculate the Pearson's Correlation Coefficient between the data, *x*, and the time-periods, *y*.

O-2.6.2. Estimating the Serial Correlation Coefficient. Consider benzene results taken from quarterly groundwater samples at well MW01 in Site A from 1998–2000. Benzene has been detected during all of these sampling events, so no proxy concentrations were derived. Also, notice how the numbers 1 through 10 replace the actual sample dates.

Time	Jan-98	Apr-98	Jul-98	Oct-98	Apr-99	Jul-99	Oct-99	Apr-00	Jul-00	Oct-00
Time Period Number	1	2	3	4	5	6	7	8	9	10
Concentration (µg/L)	12.2	3.79	3.42	5.47	0.81	1.84	7.56	4.3	2.68	6.17

O-2.6.2.1. For the concentration (*X*),

$$\sum_{i=1}^{10} x_i = 48.24, \sum_{i}^{10} x_i^2 = 329.8, \bar{x} = 4.824, \text{ and } s_x = 3.284.$$

O-2.6.2.2. For the time period (*Y*),

$$\sum_{i=1}^{10} y_i = 55, \sum_{i}^{10} y_i^2 = 385, \bar{y} = 5.5, \text{ and } s_y = 3.028.$$

The cross term is:

$$\sum_{i=1}^{10} x_i \, y_i = 240.2 .$$

O-2.6.2.3. Using Paragraph O-2.4.1, we see that the Pearson correlation coefficient, *r*, between the concentration (*X*) and the time period (*Y*) gives a serial correlation coefficient of:

$$r = \frac{240.2 - (10 \times 4.824 \times 5.5)}{9 \times 3.284 \times 3.028} = -0.2813 .$$

O-2.7. Kendall's Coefficient of Rank Correlation. In instances where data do not follow a normal or other known distribution, it is still possible to test for the significance of association between two variables. Kendall's coefficient of rank correlation, also referred to as Kendall's τ (the Greek letter tau), is a measure of correlation that may be used for variables that are at least ordinal in nature (i.e., variables with values that can be ranked). It is frequently encountered in ecological applications such as counting of fish species in a stream in different seasons.

O-2.7.1. Introduction. Kendall's τ does not assume any particular data distribution and accommodates censored values. Non-detected results should be assigned a value smaller than the lowest measured value. As the test depends only upon signs of the differences between data points (or the ranks), information about magnitudes of these differences is not used; as a result, the test possesses less power than its parametric counterpart, Pearson's *r* (i.e., a larger number of data points are required to identify a correlation using Kendall's τ). However, Kendall's τ is advantageous because assumptions about the underlying data distribution are not required, and it is less sensitive to outliers and censored values than a parametric test.

O-2.7.1.1. Kendall's τ is also invariant with respect to monotonic transformations of the variables. For example, the calculated value of τ will be identical to the calculated value for log-transformed variables. See the discussion at the end of Paragraph O-2.5 for more details. It should also be noted that for the same data, the value for Kendall's τ is generally lower than for Spearman's *r* (Conover, 1980). However, statistical tests for $\gamma = 0$ are generally in agreement between the two.

O-2.7.1.2. Kendall's τ for small sample sizes is appropriate for data with fewer than 40 samples (Gilbert, 1987); the EPA suggests using this method with data sets fewer than 10 samples. Tied observations (when two or more measurements are equal) degrade the statistical power and should be avoided, if possible, by recording the data to sufficient accuracy. If the number of samples becomes too large, the calculations become cumbersome to do by hand. Directions for calculating Kendall's τ for a small sample size (less than 10 samples) are presented in Paragraph O-2.7.2 and an example is presented in Paragraph O-2.7.3. Extensions of Kendall's τ for larger sample sizes are explained with the Mann-Kendall test for trends in Appendix P. In that Appendix, the time variable corresponds to the *X* variable here, and the *X* variable in Appendix P corresponds to the *Y* variable here.

O-2.7.2. Directions for Kendall's Coefficient of Rank Correlation. Let $(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)$ represent pairs of measurements of variables *X* and *Y*. Order the pairs from least to greatest by the x value $(x_{(1)},\ y_{x_{(1)}}),(x_{(2)},\ y_{x_{(2)}}),\ldots,(x_{(n)},\ y_{x_{(n)}})$. Here the notation $y_{x_{(i)}}$ indicates the *Y* measurement that corresponds to the *i*th *X* measurement ordered from least to greatest. The test statistic *S* is then calculated:

$$S = S^+ - S^-$$

where S^+ is the number of positive ("concordant") pairs: $(y_{x_{(i)}},\ y_{x_{(j)}})$ with $i < j$ and $y_{x_{(i)}} < y_{x_{(j)}}$. Likewise, S^- is the number of negative ("discordant") pairs: $(y_{x_{(i)}},\ y_{x_{(j)}})$ with $i < j$ and

$$y_{x_{(i)}} > y_{x_{(j)}}.$$

It can be shown that there are a total of $n(n-1)/2$ possible pairwise comparisons for a set of *n* pairs $(y_{x_{(i)}},\ y_{x_{(j)}})$. The sample statistic Kendal1's τ, is:

$$\tau = \frac{S}{n(n-1)/2}\ .$$

Note that differences of zero are not included in the test statistic (and should be avoided, if possible, by recording data to sufficient accuracy). However, an adjustment for ties may be made by calculating Kendall's τ_b ("tau b")

$$\tau_b = \frac{S}{\sqrt{\left(\frac{n(n-1)}{2} - n'_X\right)\left(\frac{n(n-1)}{2} - n'_Y\right)}}\ .$$

The quantities n'_X and n'_Y denote the number of ties for the *X* variable and *Y* variable, respectively. In particular, if there are *n* pairs of values (x_i, y_j), so that the measured values of *X* are $x_1, x_2, \ldots x_n$, then n'_X is the number of pairs (x_i, x_j), where *i* > *j*, for which $(x_i - x_j) = 0$ or for which this difference cannot be determined to be either positive or negative because of data censoring. For example, assume that there are multiple censoring limits for non-detects (e.g., < 3 and < 5), and *X* is the set of *n* = 5 values {< 1, < 3, < 5, 2, 10} with the corresponding *Y* values {2, 4, 5, 7, 9}, so that, for example, the first pair of results (x_1, y_1) is $(<1,\ 2)$. There are five tied pairs for the measured values of *X*: (< 1, < 3), (< 1, < 5), (< 3, < 5), (< 3, 2), and (< 5, < 2). Therefore, $n'_X = 5$. As there are no tied values for Y, $n'_Y = 0$. Note that when $n'_X = n'_Y = 0, \tau_b = \tau$. Tied values tend produce larger values for τ_b relative to the corresponding values for τ.

O-2.7.2.1. Table O-1 presents the resulting matrix of differences when applying the steps above. Fill in the blank spaces with a 1 if the value at the top of the column exceeds the value at the left of the row. Fill in 0 if they are equal, and fill in –1 otherwise. Then sum the values across rows and add up the sums to get *S*.

Table O-1.
Matrix of Differences for Kendall's τ

***Y* Measurements**	$y_{x_{(2)}}$	$y_{x_{(3)}}$	…	$y_{x_{(n)}}$	**Sum of Row**
$y_{x_{(1)}}$					
$y_{x_{(2)}}$					
….					
$y_{x_{(n-1)}}$					
					S

O-2.7.2.2. Use Table B-10 of Appendix B to determine the probability (*p*) using the sample size (*n*) and the absolute value of the statistic *S* if $n \le 10$.

O-2.7.2.3. For testing H_0: $\gamma = 0$ against H_A: $\gamma \ne 0$ at significance level α, reject H_0 if $p < \alpha/2$.

O-2.7.3. Example of Kendall's Rank Correlation Coefficient. Consider the same data set presented in Paragraphs O-2.4.2 and O-2.5.2 for chromium and lead in subsurface soil background (in mg/kg). Although these data are for continuous variables, it is possible to determine the rank correlation between chromium and lead using Kendall's τ.

O-2.7.3.1. First the data must be ordered by the chromium measurements as shown below.

Sample	**Chromium**	**Lead**
EPC-BG08	3.84	3.35
EPC-BG03	4.26	4.19
EPC-BG05	4.53	3.66
EPC-BG01	4.60	3.50
EPC-BG04	5.28	3.91
EPC-BG02	5.29	4.16
EPC-BG06	5.74	4.31
EPC-BG07	5.86	4.19

O-2.7.3.2. Then, create Table O-2 for the lead measurements as described in Paragraph O-2.7.2.

O-2.7.3.3. From Table O-2, S = 15. There are n = 8 pairs of lead and chromium measurements. Therefore, Kendall's tau is:

$$\tau = \frac{S}{n(n-1)/2} = \frac{15}{8(8-1)/2} = 0.536\,.$$

As there is one tie for the lead measurements (two measurements equal 4.19)

$$\tau_b = \frac{S}{\sqrt{\left(\frac{n(n-1)}{2} - n'_X\right)\left(\frac{n(n-1)}{2} - n'_Y\right)}} = \frac{15}{\sqrt{\left(\frac{8(8-1)}{2} - 0\right)\left(\frac{8(8-1)}{2} - 1\right)}} = 0.546.$$

O-2.7.3.4. To test whether the population correlation coefficient differs from 0 with 90% confidence (α = 0.05), look up the value of p corresponding to S = 15 for n = 8 in Table B-10. Owing to the tied value for lead, S = 15 does not appear in the table. Ideally, the data should have been recorded with more accuracy to break the tie. In this case, the value for S = 14 will be used to give p = 0.054 > α/2 = 0.05. We conclude that the population correlation coefficient does not differ significantly from zero with 90% confidence although further study may be needed.

O-2.8. Covariance. A statistic related to the correlation coefficient is covariance. Covariance is a measure of the linear association between two random variables, X and Y. If covariance is positive, large values of X tend to be associated with large values of Y and vice versa. If covariance is negative, large values of X tend to be associated with small values of Y and vice versa. The sample covariance is calculated as

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^{n} x_i\, y_i - n\,\bar{x}\,\bar{y}}{(n-1)}\,.$$

O-2.8.1. Pearson's correlation coefficient is derived from the covariance by dividing covariance by the sample standard deviations of X and Y.

O-2.8.2. Covariance is rarely used because the magnitude of its value is difficult to interpret. In particular, changes in scale cause changes to the covariance; that is, covariance is not invariant to changes in scale. For example, if X is multiplied by 100, its covariance with Y will also go up by a factor of 100, while its correlation with Y will remain the same.

Table O-2.
Matrix of Differences for Kendall's τ for Data in O-2.7.3.1

Lead Measurements	4.19	3.66	3.50	3.91	4.16	4.31	4.19	**Sum of Row**
$y_{x_{(1)}} = 3.35$	1	1	1	1	1	1	1	7
$y_{x_{(2)}} = 4.19$		–1	–1	–1	–1	1	0	–3
$y_{x_{(3)}} = 3.66$			–1	1	1	1	1	3
$y_{x_{(4)}} = 3.50$				1	1	1	1	4
$y_{x_{(5)}} = 3.91$					1	1	1	3
$y_{x_{(6)}} = 4.16$						1	1	2
$y_{x_{(7)}} = 4.31$							–1	–1
								S = *15*

APPENDIX P

Comparing Laboratory and Field Data

P-1. Introduction. Interpreting field data may arise in the SI and RI phases of a CERCLA project. The following discussion applies to comparing field data results to laboratory results.

P-1.1. As previously discussed, there is an inherent relationship among variability, the statistical decision confidence required, and the number of data points one must have to make the decision. There is a trade-off between cost and data quality (level of confidence for the decision-making). In general, cost and the level of confidence increase as the number of samples increases. In fact, a small set of very high quality individual measurements (e.g., from a fixed-laboratory analytical method) is frequently not as desirable as a large number of lower quality measurements (e.g., from a field analytical method). If rapid and inexpensive methods of sampling and analysis were available for the SI, a larger number of samples could be used to characterize the study area, reducing both cost and decision uncertainty. However, such methods with sufficient reliability are not always available.

P-1.2. There are many innovative field-based sampling and analysis techniques and technologies available to environmental scientists. Because of the ability to reproduce these sampling techniques with an acceptable level of accuracy and at relatively low cost, investigators can still make decisions with confidence based on field analyses.

P-1.3. When applying field analytical technologies to a given site, the project team often collects larger sample aliquots for a percentage of the field samples to ensure that the field methods are providing reasonably precise, accurate, and representative results. Each aliquot is thoroughly homogenized (i.e., unless VOCs are being analyzed) and split into a pair of duplicate samples; one sample is analyzed by the field method and the remaining sample of the duplicate pair is sent to a fixed laboratory for analysis. The results of the laboratory and field analyses are then compared to assess the usability of the field results.

P-1.4. Although the EPA has generally specified splitting 10% of screening samples with a fixed laboratory for confirmation analysis, this is an arbitrary criterion. Furthermore, there is little guidance on how to compare field and fixed laboratory results and the criteria for acceptable agreement. Therefore, a number of possible approaches are available and discussed here, including the following.

P-1.4.1. Relative percent difference (RPD).

P-1.4.2. Correlation analysis.

P-1.4.3. Regression analysis.

P-1.4.4. Group comparisons.

P-1.4.5. Percent decision match.

P-1.5. Project planners should be sensitive to the possible comparison methods so that sampling design is appropriate for the data collected and the decision to be made at their particular site.

P-2. Relative Percent Difference. The RPD for a duplicate pair of measurements (x_1, x_2) is the absolute value of the difference between the measurements divided by the mean of the measurements $\bar{x}$, expressed as a percentage:

$$RPD = \frac{|x_1 - x_2|}{\bar{x}} \times 100 \ .$$

P-2.1. The RPD is simple to calculate and has historically been used to compare two sets of data. The field values and the corresponding laboratory values are treated as duplicate pairs, and an RPD is calculated for each pair. It should be noted that, as it is usually used for environmental applications, the RPD is not a statistically based measure of agreement. The approach is semi-quantitative at best, and, in general, is not recommended. Acceptance limits for the RPDs tend to be arbitrarily defined and unrelated to acceptable tolerances for uncertainty (i.e., the RPD acceptance limits are not derived from statistically based data quality objectives for the project). Furthermore, the EPA has not established fixed acceptance limits for the RPDs of field duplicates, though EPA Region II has specified field duplicate acceptance limits for metals for data review.

P-2.2. The RPD limit for field duplicates is 50% for water and 100% for soils. RPD values from intra-laboratory studies are available for most SW-846 methods, but the values represent only the analytical component of the variability. As the RPD is proportional to the absolute difference, it is not useful for evaluating bias. Moreover, in terms of project decision-making, a process has not been developed to readily quantify the uncertainty associated with field results, nor has a range of acceptable RPD results been developed to determine whether field results are within decision limits.

P-3. Correlation Analysis**.**

P-3.1. Field data can be compared to confirmation data, typically fixed laboratory data, using correlation analysis. In this case, the data are paired and plotted on a graph, and a Pearson's *r*,[*] which is a measure of the degree of linear association between the two sets of data, is calculated. Paired statistical tests are useful because they can be used to determine whether a screening-level method is producing data that are significantly different from a definitive method. Higher values of Pearson's *r* are preferred, as this indicates increasing similarity between the field and confirmation data. For sufficiently high values of Pearson's *r*, the field data can reliably be used as a proxy for the confirmation data. As previously stated,

[*] Appendices O and Q.

there are no fixed limits for comparison, but Appendix O provides some guidance for assessing correlation results in terms of values of Pearson's *r*.

P-3.2. However, there are a number of problems with using correlation analysis as a comparison tool. A principal problem is that correlation does not imply a cause-and-effect type of relationship or provide predictive capabilities. In other words, correlation analysis cannot be relied upon to show how variable *X* affects variable *Y*, or how *X* is a predictor of unknown values of *Y*. Thus, correlation analysis is intended as a statistical tool to simply show how two variables are linearly related and the strength of this relationship. An additional problem, or complexity, with correlation analysis is that the principal statistic reported in the analysis, Pearson's *r*, requires the *X* and *Y* variables to possess a bivariate normal distribution* (not only must *X* and *Y* be normal but the "joint variation" must also be normal; that is, if every possible (*x*, y) pair were available, *Y* must be normal for every fixed value *X* = *x* and *X* must be normal for every fixed value *Y* = *y*). Finally, it is entirely possible that data sets paired in order of concentration will show linear correlation when the absolute differences between them are very large, but in some manner proportional. Thus, along with other measures, if the data give a good linear or curvilinear fit with strong correlation, this may be taken to support but not prove confirmation between results.

P-4. Regression Analysis. Field data are often compared to confirmation data, typically fixed laboratory data, using regression analysis. In this case, the data are paired and plotted on a graph and a best-fit line is created. The regression model can provide information regarding the magnitude of the difference or the functional relationship between the screening-level and definitive methods, so that screening-level data can be converted to definitive data.

P-4.1. However, functional relationships between screening-level and definitive data are often inappropriately established. Classical linear regression analysis, as presented in Appendix O, is not appropriate for this analysis because both screening-level data (the "dependent" variable) and laboratory concentrations (the "independent" variable) are measured values, and because the laboratory concentrations (the "independent" variable) has more than a negligible amount of variability. For example, the laboratory concentrations could be selected as the "independent" variable *X* to generate a regression line of the form,

$$y = b_1 x + b_0 .$$

P-4.2. This implies

$$x = (1/b_1)y + (-b_0/b_1) .$$

P-4.3. However, the alternative selection of *Y* as the "independent" variable would produce a regression line,

* Appendix O.

$$x = b_1' \ y + b_0' \ .$$

P-4.4. Unfortunately, $b_1' \neq 1/b_1$ and $b_0' \neq (-b_0 / b_1)$. In other words, the classic or ordinary least squares (OLS) line produced from *X* and *Y* measurement data depends upon whether *X* or *Y* is arbitrarily selected as the independent variable. Therefore, it would be inappropriate to generate a regression line to "convert" screening level measurements to laboratory concentrations (or vice versa).

P-4.5. In place of OLS linear regression, reduced major axis (RMA) regression is a reasonable parametric approach, while the Kendall-Theil line is a desirable non-parametric approach for establishing a linear relationship. Advantages to reduced major axis regression are the following.

P-4.5.1. While a classic (OLS) regression line of the form $y = b_1x + b_0$ minimizes the sum of the distances in the *y*-direction from the regression line to each observed point y_i, the RMA line minimizes error for both *X* and *Y* by minimizing the sum of the areas of right triangles formed by horizontal and vertical lines extending from each observation (x_i, y_i) to the best-fit straight line (Helsel and Hirsch, 1992, p. 276).

P-4.5.2. Unlike OLS regression, RMA regression produces a unique line regardless of which variable, *X* or *Y*, is used as the response or independent variable.

P-4.6. RMA regression is used to model the correct functional relationship between two variables when both variables possess comparable measurement error. It is commonly used to evaluate biological data. All of the assumptions required for OLS regression are required for RMA regression (e.g., the residuals must be normally distributed). RMA regression has also been called "line of organic correlation," "geometric mean functional regression," and "Maintenance of Variance-Extension" (Helsel and Hirsch, 2003). Reduced major axis regression should not be confused with an alternative approach referred to as "major" or "principal axis" regression. Major axis regression is often used in lieu of RMA regression as it is conceptually similar; the best fit line minimizes the sum of the squares of the perpendicular distances between the line and each plotted observation (rather than the areas of right triangles). Both reduced major axis and major axis regression are often referred to as "model II" regression (OLS regression is "model-I" regression).

P-4.7. The slope (b_1'') and intercept (b_0'') of the RMA regression line $y = b_1''x + b_0''$ are as follows:

$$b_1'' = sign[r](s_y / s_x)$$

$$b_0'' = \bar{y} - b_1'' \bar{x}$$

where $sign[r]$ is the algebraic sign of Pearson's *r*; s_y and s_x are the sample standard deviations of Y and X, respectively; and $\bar{y}$ and $\bar{x}$ are the sample arithmetic averages of Y and X, respectively. Like an OLS regression line, the RMA regression line passes through the point $(\bar{x}, \bar{y})$, but (unlike an OLS regression line) the slope does not depend upon the magnitude of the regression coefficient *r*. Given the OLS regression lines $y = b_1 x + b_0$ and $x = b_1' y + b_0'$, an alternative expression for the major axis regression slope is:

$$b_1'' = sign[r]\sqrt{b_1 / b_1'} \ .$$

P-4.8. Thus, the slope of the RMA regression line is essentially the geometric mean of the OLS slopes b_1 and $1/b_1'$ (hence the use of the terminology "geometric mean regression"). An equivalent expression for the RMA slope is:

$$b_1'' = b_1 / r$$

Note that, because $r \leq 1$, the RMA slope will be equal to or greater than the slope of the corresponding OLS regression line.

P-4.9. Confidence limits can be calculated for the slope and intercept of the RMA regression line. The $(1 - \alpha)100\%$ confidence interval for the slope is as follows (Warton, 2005)

$$\left[b_1''\left(\sqrt{B+1} - \sqrt{B}\right), b_1''\left(\sqrt{B+1} + \sqrt{B}\right)\right] \tag{P-1}$$

where

$$B = \frac{F_{1-\alpha,1,n-2}\,(1 - r^2)}{n-2}$$

$F_{1-\alpha,1,n-2}$ is the critical value of the *F*-distribution with 1 degree of freedom in the numerator and $n - 2$ degrees of freedom in the denominator. The confidence limits for the intercept are:

$$b_0'' \pm t_{1-\alpha/2,n-2}\, s_0 \ . \tag{P-2}$$

P-4.10. The quantity s_0 denotes the estimated standard deviation of the intercept of the OLS regression line $y = b_1 x + b_0$, which may be determined from the equation:

$$s_0 = \sqrt{\frac{s^2}{n} + \bar{x}\, s_1^2} \ .$$

P-4.11. The quantity s^2 denotes the estimated variance of residuals of the OLS regression line $y = b_1 x + b_0$ and s_1^2 the estimated variance of slope of the OLS slope

$$s^2 = \frac{s_y^2\,(n-1)\left(1-r^2\right)}{(n-2)}$$

$$s_1^2 = \frac{s^2}{s_x^2\,(n-1)}$$

where

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{(n-1)}$$

and

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{(n-1)}\ .$$

P-4.12. The reader is referred to software that can be used to calculate RMA regression lines as well as confidence limits for the slopes and intercepts (Bohonak, 2004), though the software does not calculate the confidence limits of the slope using Equation P-1 but using an approximation that produces a similar result:

$$b_1'' \pm t_{1-\alpha/2,\,n-2}\, s_1\ .$$

P-4.13. A non-parametric approach for establishing a linear relationship is the Kendall-Theil line. The line takes the form: $y = \hat{b}_1 x + \hat{b}_0$. The slope $(\hat{b}_1)$ is computed by comparing each data pair to all others in a pairwise fashion. A data set of *n* (*x*, *y*) pairs will result in *n*(*n* – 1)/2 pairwise comparisons. For each of these comparisons, a slope is computed by

$$m_{ij} = \frac{\left(y_j - y_i\right)}{\left(x_j - x_i\right)} \text{ for all } i < j\,;\ i = 1, 2, \ldots, (n-1)\,; \text{ and } j = 2, 3, \ldots, n\,.$$

P-4.14. Note that m_{ij} is the value of the random variable, M . The slope $(\hat{b}_1)$ and intercept $(\hat{b}_0)$ are estimated as follows:

$$\hat{b}_1 = \tilde{m}\text{, where } \tilde{\mathrm{m}} \text{ is the median of } M$$

and

$$\hat{b}_0 = \tilde{\mathrm{y}}\text{ - }\hat{b}_1\tilde{x}\text{, where } \tilde{\mathrm{y}} \text{ and } \tilde{\mathrm{x}} \text{ are the medians of } Y \text{ and } X\text{, respectively.}$$

P-4.15. Therefore, the line passes through the point $(\tilde{x}, \tilde{\mathrm{y}})$, analogous to the ordinary least squares regression line, which passes through the point $(\overline{x},\ \overline{y})$. The Kendall-Theil line is closely related to the Kendall's τ (see Appendix O) because the hypothesis test that $\hat{b}_1$ is equal to zero is the same as the hypothesis test that τ is equal to zero. The Kendall-Theil line has the desirable property of a nonparametric estimator: it is almost as efficient as the parametric estimator when all assumptions of normality are met, and is much better when those assumptions are not met (Helsel and Hirsch, 2003). A confidence limit for the slope of the line can be calculated by ordering the slopes m_{ij} for all $i < j$; $i = 1, 2, \ldots, (n-1)$ and $j = 2, 3, \ldots, n$ from smallest to largest, and selecting the r^{th} and s^{th} slopes such that the following inequality holds true:

$$P\left(m_{(r)} < M < m_{(s)}\right) \geq 1 - \alpha.$$

P-4.16. For more details about this confidence limit, see "Statistical Methods in Water Resources" (Helsel and Hirsch, 2003) or "Practical Nonparametric Statistics" (Conover, 1980).

P-5. Group Comparisons. In a manner similar to the comparison between background and on-site data, screening and definitive confirmation data can be compared as groups. After verifying that the minimum assumptions of the various tests are met, group means and variances can be compared using *t*- and *F*-tests or their non-parametric equivalents (See Appendices M and N). In this case, the project team must decide on the decision confidence required, most likely α will be 0.2 or less. Methods for determining decision confidence levels are discussed in Appendix K.

P-5.1. The following provides a review of issues that must be considered when applying the method of group comparisons; the review primarily focuses on comparing distinctly different groups of data. Consider a site that contains areas of both high and low contamination. Given the extreme divergence in contamination levels, there will be different population means across the sampled areas. Sample data analyzed using Field Method A cannot simply be compared to the entire set of sample data using Laboratory Method B with a two-sample *t*-test (refer to Appendix N) because of the different mean levels of the measured contaminant. For this approach to be viable (i.e., two sample *t*-test based on field and laboratory methods), the underlying population would need to be relatively homogeneous. If this condition is not met, statistical tests for paired data would need to be used.

P-5.2. Paired statistical tests are recommended to determine whether Field Method A and Laboratory Method B are significantly different. To conduct these tests, an aliquot is homogenized and split into duplicates (it is possible the sample extracts would be split as well). One duplicate is analyzed by Method A and the other analyzed by Method B. For each data pair, the researcher evaluates the difference in results provided between Methods A and B. If the results from Method A are not different from corresponding results provided by Method B and the differences are normally distributed, then on the average, the difference between the two methods is zero. However, it should be noted that, as the differences are usually calculated over a range of concentrations (rather than at a single concentration), an average difference of zero does not necessarily demonstrate that Methods A and B are comparable. For example, it would be possible for Method A to produce much smaller values than Method B at low concentrations but much larger values at high concentrations so that, on the average, the differences between Method A and Method B over the entire concentration range is nearly zero. If Methods A and B are different, then the researcher should establish a functional relationship ($X_B = f(X_A)$) using regression analysis to "convert" the Field Method A results (X_A) to the corresponding laboratory Method B results (X_B) (see Paragraph P-4 for a discussion of regression analysis). The computed relationship, though, would need to quantify the uncertainty associated with the conversion. If this uncertainty is small relative to the uncertainty contributed by the field component, then the conversion uncertainty can be ignored and the "converted results" (X_B) used directly (i.e., can be treated as if they were directly obtained from a definitive laboratory method).

P-6. Percent Decision Match (PDM). The PDM may be a practical and useful approach to confirmation testing. The PDM is a qualitative evaluation strategy, as opposed to a more traditional statistical or quantitative strategy. For example, in the PDM, the decision error is not quantified and the variability in PDM results for a study area is not incorporated into the analysis. The PDM approach may be useful certain data quality objectives, namely to determine whether site contamination exceeds a specified decision limit.

P-6.1. The PDM is calculated as the number of times both data points in a data pair lead to the same conclusion divided by the total number of data pairs, expressed as a percentage:

$$\text{PDM} = \frac{\mathit{Number\ of\ Decision\ Matches}}{\mathit{Number\ of\ Data\ Pairs}}.$$

P-6.1.2. For example, suppose the regulatory threshold to which the data will be compared is fixed at 100 ppm. Suppose further that 100% of the data points from the screening technology are less than the threshold and the mean concentration is 50 ppm. Now, let us suppose that the definitive method of analysis systematically produces lower results and the mean concentration is 10 ppm. If both the screening data and the definitive data lead to the same conclusion, namely, that all of the samples are less than the threshold, is the difference between the absolute values of the screening and definitive analyses of any real significance? A PDM greater than 90% has historically been found to be acceptable to regulators in a number of differing jurisdictions.

APPENDIX Q

Trend Analysis

Q-1. Introduction.

Q-1.1. This Appendix presents tools for detecting and estimating trends in environmental data. Trends may be spatial or temporal and can take various forms, including steady increases or decreases or a steep increase or decrease at a point in time or space. Detecting and estimating 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. The detection of trends may be complicated by the overlaying of long- and short-term trends, cyclical effects such as seasonal or weekly systematic variations, autocorrelations, or impulses or jumps from interventions or procedural changes. Trend is just one of several aspects of time series, the study of data with respect to time. Time series consists of trends, seasonal variation or seasonality, cyclical variation or repetitive trends, and irregular activity (Kvanli et al., 1996).

Q-1.2. The following subparagraphs present methods for detecting seasonal or temporal repetitive trends, correcting for seasonality, and testing procedures for trends using regression techniques and more robust trend estimation procedures. The investigations of trends in this Appendix are limited to one-dimensional domains, trends in a constituent concentration over time. This Appendix does not address spatial trends (with two- and three-dimensional domains) and trends over space and time (with three- and four-dimensional domains), which may involve sophisticated geostatistical techniques such as kriging (Appendix R). Gilbert (1987) and Gibbons (1994) provide additional resources for trend analysis.

Q-2. Identifying Seasonality and Other Repetitive Trends. Seasonality is one factor that accounts for changes in concentrations over time. Environmental monitoring data are likely to exhibit seasonality. According to Kvanli, et al. (1996), seasonality is a predictable, periodic increase or decrease that occurs within a time period or cycle, such as 1 year. The key to identifying such trends is the repetition of the same pattern for each cycle. Identifying seasonality or other repetitive trends (i.e., persistent cyclic variations) is necessary before long-term increasing or decreasing temporal trends can be evaluated in environmental data. To identify these, a project team should visually inspect plots of data across time for seasonal or repetitive trends. Project teams should justify all seasonal trends identified visually with respect to site history, geology, chemistry, and professional judgment.

Q-2.1. Overview of Seasonal Trends.

Q-2.1.1. Generally, seasonality is not the primary focus of evaluating monitoring data for temporal trends. As such, data should be adjusted to remove the seasonal effects so that other temporal trends may be studied. For instance, if groundwater concentrations are diluted every spring by high recharge, true changes in groundwater may be masked by this effect. Likewise, if low water flow in fall leads to higher concentrations in groundwater that do not represent more leaching from a source area, then these effects should be accounted for in data evaluation. Seasonal effects may be removed by adjusting the sample data or using statistical methods unaffected by such relations. Adjustments to the sample data are described in this Paragraph. The subsequent Paragraph provides details about statistical tests that account for data with seasonal variability.

Q-2.1.2. There are various methods to de-seasonalize data. If the seasonal pattern is regular, it may be modeled with a sine or cosine function. Moving averages can be used, or differences (of order 12 for monthly data, for example) can be used. However, time series models may include rather complicated methods for de-seasonalizing the data. A simpler method is presented in EPA 530-SW-89-026 for applications to any seasonal cycle. For environmental data, seasonal cycles typically occur annually, monthly, or quarterly. Directions for the EPA method are presented in Paragraph Q-2.2, followed by an example in Paragraph Q-2.3. Although EPA’s method assigns seasonality as a monthly cycle, this method can be applied with other seasonal or repetitive cycles by replacing “monthly” with the appropriate cycle.

Q-2.2. Directions for Correcting Seasonality in Data. To correct seasonality with time series data, directions are provided for monthly data that demonstrate a yearly cycle.

Q-2.2.1. Assume *n* years of monthly data are available.

Q-2.2.2. Let x_{ij} denote the unadjusted observation for the i^{th} month and the j^{th} year.

Q-2.2.3. Compute the average concentration for month *i* over the *n*-year period:

$$\bar{x}_i = \frac{(x_{i1} + ... + x_{in})}{n}.$$

This average represents the average of all observations taken in different years, but during the same month.

Q-2.2.4. Calculate the grand mean, $\bar{x}$, of all 12 *n* observations:

$$\bar{x} = \sum_{i=1}^{12} \frac{\bar{x}_i}{12}.$$

Q-2.2.5. Compute the adjusted concentrations,

$y_{ij} = x_{ij} - \bar{x}_i + \bar{x}$.

Q-2.2.6. The difference $x_{ij} - \bar{x}_i$ removes the average effect of month *i* from the monthly data. The grand mean ($\bar{x}$) must be added (on the right hand side of the equation) so that the mean of the adjusted y_{ij} values, $\bar{y}$, is equal to the grand mean ($\bar{x}$) of the unadjusted values.

Q-2.3. Correcting Seasonality with Time Series Data (Based on Monthly Data with a Yearly Cycle*)*. Consider evaluating seasonality for the monthly average temperature (in degrees Fahrenheit) in Austin, Texas, from 1995 through 1998 (Table Q-1). A time plot of the data is presented in Figure Q-1.

Table Q-1.
Monthly Average Temperature (°F) in Austin, Texas, from 1995 through 1998

Month-Year	Temper-ature		Month-Year	Tempera-ture		Month-Year	Tempera-ture		Month-Year	Tempera-ture
Jan-95	50.03		Jan-96	47.10		Jan-97	46.00		Jan-98	53.06
Feb-95	53.00		Feb-96	53.38		Feb-97	50.15		Feb-98	52.21
Mar-95	57.00		Mar-96	52.84		Mar-97	60.68		Mar-98	55.90
Apr-95	62.23		Apr-96	62.77		Apr-97	59.57		Apr-98	62.70
May-95	71.94		May-96	73.67		May-97	67.87		May-98	73.68
Jun-95	74.23		Jun-96	77.13		Jun-97	74.97		Jun-98	79.60
Jul-95	79.26		Jul-96	81.06		Jul-97	78.45		Jul-98	82.10
Aug-95	78.45		Aug-96	77.42		Aug-97	77.94		Aug-98	80.19
Sep-95	74.07		Sep-96	72.93		Sep-97	75.03		Sep-98	78.73
Oct-95	66.06		Oct-96	66.13		Oct-97	65.84		Oct-98	68.10
Nov-95	55.77		Nov-96	56.55		Nov-97	53.83		Nov-98	60.37
Dec-95	51.37		Dec-96	51.93		Dec-97	47.50		Dec-98	49.81

THIS SPACE INTENTIONALLY LEFT BLANK

Q-2.3.1. The plot indicates the seasonality plays a role in this data. There are *n*=4 years of monthly data. The average temperature for each month and the grand average for all months are presented below.

Month	**Average Temperature**
January	49.05
February	52.19
March	56.61
April	61.82
May	71.79
June	76.48
July	80.22
August	78.50
September	75.19
October	66.53
November	56.63
December	50.15
Grand Average	64.60

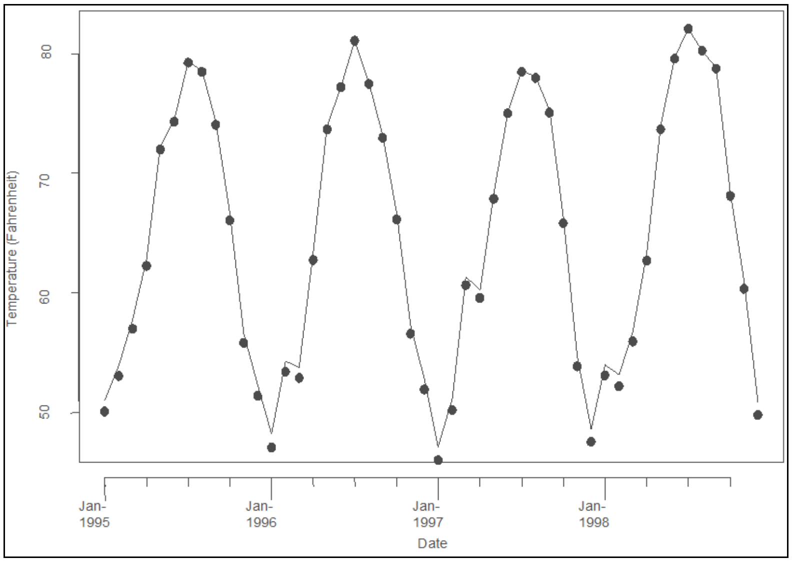


Figure Q-1. Monthly Average Temperature (°F) in Austin, Texas, from 1995 through 1998.

Q-2.3.2. The average January temperature is simply the average of all the January temperatures, no matter the year:

$$\bar{x}_{January} = \frac{50.03 + 47.10 + 46.00 + 53.06}{4} = 49.05\,.$$

Q-2.3.3. The other monthly averages are estimated in the same fashion. The grand average is simply the average of all of the monthly averages:

$$\bar{x} = \frac{49.05 + 52.19 + 56.61 + 61.82 + 71.79 + 76.48 + 80.22 + 78.50 + 75.19 + 66.53 + 56.63 + 50.15}{12} = 64.60$$

Q-2.3.4. The adjusted averages are presented in Table Q-2. The adjusted Jan-1995 temperature, for example, was estimated by the following: adjusted temperature = 50.03 – 49.05 + 64.60 = 65.58. Figure Q-2 is a plot of the adjusted temperatures. The vertical scale of the plot is the same as the plot of the adjusted data to emphasize that the seasonal variation has been smoothed out.

Table Q-2.
Adjusted Monthly Average Temperature (°F) in Austin, Texas, from 1995 through 1998

Month-Year	Temperature	Monthly average temperature	Grand average temperature	Adjusted temperature	Month-Year	Temperature	Monthly average temperature	Grand average temperature	Adjusted temperature
Jan-95	50.03	49.05	64.60	65.58	Jan-97	46.00	49.05	64.60	61.55
Feb-95	53.00	52.19	64.60	65.41	Feb-97	50.15	52.19	64.60	62.56
Mar-95	57.00	56.60	64.60	64.99	Mar-97	60.68	56.60	64.60	68.67
Apr-95	62.23	61.82	64.60	65.01	Apr-97	59.57	61.82	64.60	62.35
May-95	71.94	71.79	64.60	64.74	May-97	67.87	71.79	64.60	60.68
Jun-95	74.23	76.48	64.60	62.35	Jun-97	74.97	76.48	64.60	63.08
Jul-95	79.26	80.22	64.60	63.64	Jul-97	78.45	80.22	64.60	62.83
Aug-95	78.45	78.50	64.60	64.55	Aug-97	77.94	78.50	64.60	64.03
Sep-95	74.07	75.19	64.60	63.47	Sep-97	75.03	75.19	64.60	64.44
Oct-95	66.06	66.53	64.60	64.13	Oct-97	65.84	66.53	64.60	63.90
Nov-95	55.77	56.63	64.60	63.73	Nov-97	53.83	56.63	64.60	61.80
Dec-95	51.37	50.15	64.60	65.81	Dec-97	47.50	50.15	64.60	61.95
Jan-96	47.10	49.05	64.60	62.64	Jan-98	53.06	49.05	64.60	68.61
Feb-96	53.38	52.19	64.60	65.79	Feb-98	52.21	52.19	64.60	64.62
Mar-96	52.84	56.60	64.60	60.83	Mar-98	55.90	56.60	64.60	63.89
Apr-96	62.77	61.82	64.60	65.55	Apr-98	62.70	61.82	64.60	65.48
May-96	73.67	71.79	64.60	66.47	May-98	73.68	71.79	64.60	66.49
Jun-96	77.13	76.48	64.60	65.25	Jun-98	79.60	76.48	64.60	67.71
Jul-96	81.06	80.22	64.60	65.44	Jul-98	82.10	80.22	64.60	66.47
Aug-96	77.42	78.50	64.60	63.52	Aug-98	80.19	78.50	64.60	66.29
Sep-96	72.93	75.19	64.60	62.34	Sep-98	78.73	75.19	64.60	68.14
Oct-96	66.13	66.53	64.60	64.20	Oct-98	68.10	66.53	64.60	66.16
Nov-96	56.55	56.63	64.60	64.52	Nov-98	60.37	56.63	64.60	68.33
Dec-96	51.93	50.15	64.60	66.37	Dec-98	49.81	50.15	64.60	64.25

Q-2.4. Summary. Corrections for seasonality should be used with great caution because they represent extrapolation into the future. There should be good scientific explanation and good empirical evidence for the seasonality before corrections are made. For instance, larger than average rainfalls for two or three Augusts in a row does not justify the belief that there will never be a drought in August, and this idea extends directly to any monitoring system. In addition, the quality (bias, robustness, and variance) of the estimates of the proper corrections must be considered even in cases in which corrections are called for. If seasonality is suspected, adjusting for seasonality may not be necessary to evaluate long-term trends when appropriate statistical methods are utilized. Such methods will be discussed in the following Paragraph.

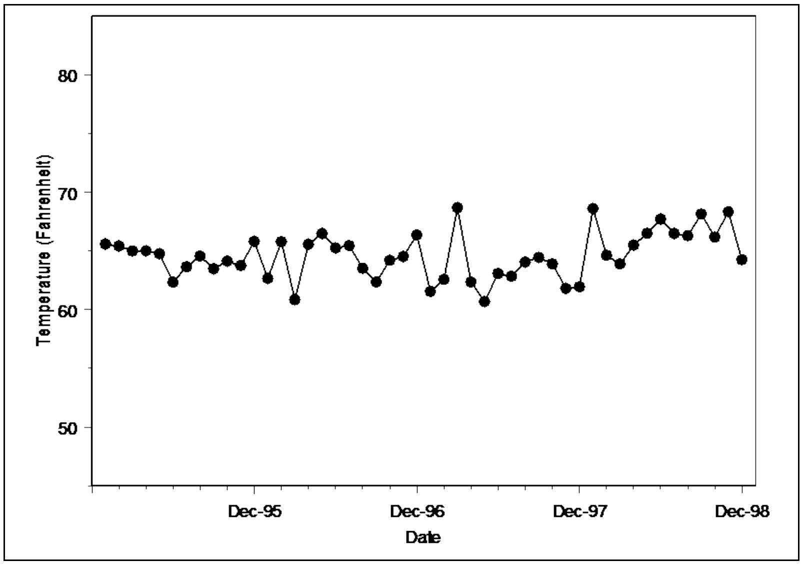


Figure Q-2. Adjusted Monthly Average Temperature (°F) in Austin, Texas, from 1995 through 1998.

Q-3. Methods for Trend Assessment.

Q-3.1. Introduction. As a first step in evaluating trends, graphical representations are recommended to identify possible trends. A plot of the data versus time is recommended for temporal data, as it may reveal long-term trends and show other major types of trends, such as cycles or impulses.

Q-3.1.1. A posting plot is recommended for spatial data to reveal spatial trends such as areas of high concentration or areas that were inaccessible. (See Appendix J for further discussion of posting plots.) Gilbert (1987) recommends smoothing time series to identify cycles and long-term trends that may be obscured by natural variation in the data. Gilbert also mentions using control charts as an effective graphical tool of trends. Control charts are presented at the end of this section.

Q-3.1.2. Most of the statistical tools presented below are applicable to environmental data; the focus is on monotonic, long-term trends (i.e., trends that are exclusively increasing or decreasing, but not both), as well as other sources of systematic variation, such as seasonality.

Q-3.1.3. There are numerous tests for trends. Trend tests, like other statistical tests, can be divided in terms of distributional assumptions. Parametric trend tests, which assume data follow a normal distribution, involve regression-based methods for estimating trends and determining if a significant trend exists. Nonparametric trend tests, which do not make assumptions about the underlying data distributions, are based on the Mann-Kendall trend test.

Q-3.1.4. Independence is crucial for parametric and nonparametric tests. The departure from independence (if data are correlated) can result in incorrect conclusions (Gibbons, 1994). To minimize the possibility that samples are not independent, Gibbons recommends a sampling frequency of no more than one sample per quarter. In practice, sampling frequency may be based on knowledge of site conditions such as groundwater flow rates.

Q-3.1.5. Regression-based methods usually are not recommended for environmental studies as a general tool for estimating and detecting trends, although they may be useful as a quick and easy-to-use screening tool for identifying strong linear trends. Regression analyses can be misleading if seasonal cycles are present, the data are not normally distributed, or the data are serially correlated (Gilbert, 1987). In such cases, Gilbert suggests that the non-parametric seasonal Kendall test is preferable to regression methods. Non-parametric trend tests are more appropriate when data do not conform to a particular distribution and when there are data below the detection limit. For groundwater monitoring, Gibbons (1994) states that non-parametric analyses are the most
reasonable estimators of trend.

Q-3.2. Regression-Based Methods. Classic procedures for assessing linear trends use regression. Linear regression is a common procedure in which calculations are performed on a data set containing pairs of observations (x_i, y_i). For temporal trends, the x_i values represent time and the y_i values represent the observations, such as contaminant concentrations. "If plots of data versus time suggest a simple linear increase or decrease over time, a linear regression of the variable against time may be fit to the data. A *t*-test may be used to test that the true slope is not different from zero (Gilbert, 1987)."

Q-3.2.1. Regression procedures are easy to apply but entail several limitations and assumptions. For example, simple linear regression (the most commonly used method) is designed to detect linear relationships between two variables; other types of regression models generally are necessary to detect non-linear relationships, such as cyclical or non-monotonic trends. Regression is also very sensitive to extreme values (outliers) and presents difficulties in handling data below the detection limit, which are commonly encountered in environmental studies.

Q-3.2.2. A regression model is of the form:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where:

Y	=	response/dependent variable
X	=	independent/explanatory variable (e.g., time)
β_0	=	“true” intercept
β_1	=	“true” slope
ε	=	random error.

Q-3.2.3. If not for the random error, ε, all of the points (x_i, y_i) would lie precisely on the line $Y = \beta_0 + \beta_1 X$. The regression model assumes that the error is a normally distributed random variable (ε) with a mean of zero and constant variance (i.e., the variance does not depend on X). In practice, β_0 and β_1 are unknown quantities and a set of n measured values (x_i, y_i) is used to estimate a regression line of the form:

$$y_i = b_0 + b_1 x_i + e_i$$

where b_0 is an estimate of β_0, b_1 is an estimate of β_1, and e_i estimates ε_i. The slope and intercept can be estimated as follows:

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

$$b_0 = \overline{y} - b_1\overline{x} \quad .$$

Q-3.2.4. The estimated “residuals” (e_i) are calculated from the equation:

$$e_i = y_i - (b_0 + b_1 x_i).$$

Q-3.2.5. Tests for normality (for example, normal probability plots as discussed in Appendix J) are required to verify the normality of the set of results $\{e_i\}$. A plot of e_i versus x_i is required to verify that the variance of the residuals is constant (i.e., not dependent upon X). Figure Q-3 shows two commonly seen residual patterns. In Figure Q-3a, the residuals show no pattern, so the assumption of constant variance is met. In Figure Q-3b, the variance of the residuals increases as the independent variable (X) increases so the assumption of constant variance is not met. Statistical software is often used to verify the normality of the residuals and constant variance because it is burdensome to do so manually. Moreover, the analyst must ensure that time plots of the data do not possess any cyclical patterns, outlier tests show no extreme data values, and data validation reports indicate that nearly all of the measurements are above detection limits.

Q-3.2.6. Because of these limitations, regression is not recommended as a general tool for estimating and detecting trends, although it may be useful as a screening tool for identifying strong linear trends. For situations in which regression methods can be applied appropriately, a solid body of literature on hypothesis testing is available that uses the concepts of statistical linear models as a basis for inferring the existence of temporal trends.

Q-3.2.7. 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; that is, if $r = 0$, the slope $b_1 = 0$ (for more details on the correlation coefficient test see Appendix O). Directions are provided in Paragraph O-2.2, followed by an example in Paragraph O-2.3.

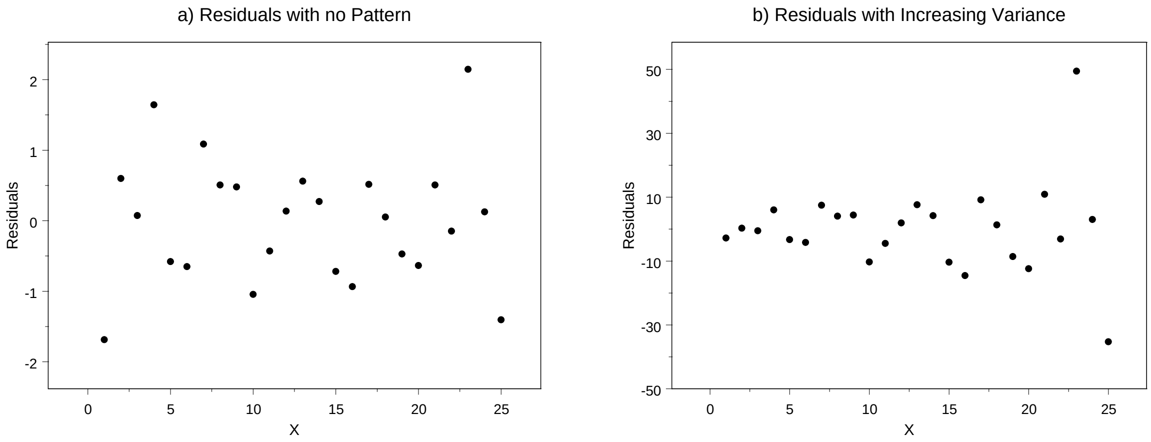


Figure Q-3. Residuals Versus the Independent Variable.

Q-3.2.8. This test assumes a linear relation between *X* and *Y* with independent, normally distributed errors and constant variance across all *X* and *Y* values. Censored values (below the detection limit) and outliers may invalidate the tests.

Q-3.2.9. If a linear trend is present, based on visual inspection or results from testing for trends, the true slope (change per unit time) may be estimated. 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 that best fits the data as a measure of strength in the trend.

Q-3.3. Non-parametric Methods.

Q-3.3.1. Introduction. Kendall's tau (Appendix O) can be used to evaluate trends. An alternative method is presented here to use for a single set of observations, x_1, x_2,..., x_n, which have been ordered by time of measurement. The test statistic S is calculated by:

$$S = S^{+} - S^{-}$$

where S^{+} is the number of pairs (x_i, x_j) with $i < j$ and $x_i < x_j$. Likewise, S^{-} is the number of pairs (x_i, x_j) with $i < j$ and $x_i > x_j$.

Q-3.3.1.1. It can be shown that there are a total of $n(n-1)/2$ possible pairwise comparisons for a set of n pairs (x_i, x_j). The sample statistic Kendal1's tau, τ, is:

$$\tau = \frac{S}{n(n-1)/2}$$

Note that differences of zero are not included in the test statistic (and should be avoided, if possible, by recording data to sufficient accuracy). However, an adjustment for ties may be made (i.e., when many ties occur), for a series of measurements x_1, x_2,..., x_n performed sequentially in time, by calculating Kendall's tau-b, τ_b:

$$\tau_b = \frac{S}{\sqrt{\left(\frac{n(n-1)}{2} - n'_X\right)\left(\frac{n(n-1)}{2}\right)}} \ .$$

The quantity n'_X is the number of tied pairs (x_i, x_j), where $j > i$, for $i = 1, 2, \ldots n$. The tie adjustment increases the magnitude of Kendall's tau and is useful for evaluating trends (or correlation) when measurements are censored.

Q-3.3.1.2. The Mann-Kendall test does not assume any particular data distribution and accommodates censored values. Non-detected results should be assigned a value smaller than the lowest measured value when the detection limit is small. Otherwise, when calculating S, pairs of results such as (3, <10), (<3, <10), and (<3, <3) should be considered to be ties and assigned a value of zero. For example, for the set of n = 4 sequential measurements {30, <10, < 20, <25}, the number of tied pairs $n'_X = 3$ for the calculation of τ_b: (<10, <20), (<10, <25), and (<20, <25). As the test only depends upon signs of differences between data points (or the ranks), information about magnitude of these differences is not used. As such, the test possesses less power than its parametric counterpart, Pearson's r (i.e., a larger number of data points are required to identify a correlation using Kendall's tau). However, Mann-Kendall is advantageous because assumptions about the underlying data distribution are not required, and it is more robust (i.e., insensitive) than a parametric test to outliers and censored values. Kendall's tau is also invariant with respect to monotonic transformations of the variable *X*. For example, the value of τ calculated for *X* will be identical to that calculated for *Ln*(*X*).

Q-3.3.1.3. Conducting the Mann-Kendall test for small sample sizes is appropriate for data with fewer than 40 samples (Gilbert, 1987); the EPA suggests using this method with data sets having fewer than 10 samples. If the number of samples becomes too large, the calculations become cumbersome by hand. Directions for the Mann-Kendall trend test for a small sample size (less than 10 samples) are presented in Paragraph
Q-3.3.2, followed by an example in Paragraph Q-3.3.3.

Q-3.3.1.4. The Mann-Kendall test is essentially a significance test under the hypothesis $\gamma = 0$ (refer to Appendix O). A trend exists if the sample statistic τ is significantly different from zero at some specified level of confidence. If there is an underlying upward trend, the differences will tend to be positive (*S* will be a large value), so a sufficiently large positive value of the sample statistic τ (e.g., a value near 1) suggests an upward trend. Conversely, if the differences tend to be negative (*S* will be a large negative value), a sufficiently large negative value of τ (e.g., a value near –1) suggests a downward trend. If the statistic τ is nearly zero (i.e., not significantly different from zero), there is no evidence of a trend. The slope of the time-ordered data plotted versus time is zero. The significance test for $\gamma = 0$ is a nonparametric test for zero slope (Gilbert, 1987). For a two-sided test the null and alternative hypotheses are:

$H_0 : \gamma = 0$: No upward or downward trend.

$H_A : \gamma \neq 0$: An upward or downward trend.

For a one-sided test

$H_0 : \gamma \leq 0$ (or $\gamma \geq 0$): No upward (or no downward trend).

$H_A : \gamma > 0$ (or $\gamma < 0$): An upward trend (or a downward trend).

Q-3.3.1.5. In practice, it is not convenient to calculate a value of τ for the data set and to compare this to a critical value of τ for the desired level of significance, τ_p (so that, for example, if $\tau > \tau_p$, there is an increasing trend at the *p*100% level of confidence). The calculations for the Mann-Kendall test are done differently for large versus small data sets. For small data sets (Paragraph Q-3.3.2), the value of *S* for the data set (rather than τ) is calculated and compared to a critical value of *S* taken from a statistical table. For large data sets, the standard normal distribution is used to determine the statistical significance of τ (Paragraph Q-3.3.4).

Q-3.3.1.6. Note that irregularly spaced measurement periods are permitted with the Mann-Kendall test (Gibbons, 1994). The test can also be modified to deal with multiple observations per time period and generalized to deal with multiple sampling locations and seasonality (Gilbert, 1987). The Mann-Kendall test for the situation in which one observation per time period is taken from one sampling location (e.g., groundwater monitoring well) is presented in Paragraph Q-3.3.2.

Q-3.3.1.7. For large sample sizes, the normal approximation to the Mann-Kendall test is used. If there are more than 10 samples, as long as there are not many tied data values, Gilbert (1987) suggests this normal approximation is quite accurate. Directions for this approximation are provided in Paragraph Q-3.3.2.4, followed by an example in Paragraph Q-3.3.2.5. Tied observations (when two or more measurements are equal) degrade the statistical power and should be avoided, if possible, by recording the data to sufficient accuracy. If the sample size is 10 or more, a normal approximation to the Mann-Kendall procedure may be used.

Q-3.3.2. Directions for the Mann-Kendall Trend Test for a Small Sample Size. List the data in the order collected over time: $x_1, x_2, \ldots, x_n$ where x_i is the datum at time t_i.

Q-3.3.2.1. Assign a proxy value to values reported as below the detection limit (DL). Note that this proxy value should be less than any measured value. Construct a Data Matrix similar to the top half of the Table Q-3.

Q-3.3.2.2. Determine the sign for each possible difference and compute the Mann-Kendall statistic, *S*, which is the number of positive signs minus the number of negative signs in the triangular table: $S = S^+$ (i.e., total number of + signs) – S^- (i.e., total number of – signs).

Q-3.3.2.3. Use Table B-10 of Appendix B to determine the probability (*p)* using the sample size (*n*) and the absolute value of the statistic *S* if $n \leq 10$.

Q-3.3.2.3.1. For testing H_0, no trend against H_A: upward trend, reject H_0 if $S > 0$ and $p < \alpha$.

Q-3.3.2.3.2. For testing H_0, no trend against H_A: downward trend, reject H_0 if $S < 0$ and $p < \alpha$.

Q-3.3.2.4. Table Q-3 presents the resulting matrix of differences when applying the steps above.

Q-3.3.2.5. The number of positive and negative differences are recorded for each row (two right most columns) and the values (within the two right most columns) are summed to obtain S^+ and S^-. Differences equal to zero are ignored.

Table Q-3.
Basic Mann-Kendall Trend Test with a Single Measurement at Each Time Point

Time x_i	t_2 x_2	t_3 x_3	t_4 x_4	**. . .** **. . .**	t_{n-1} $x_{n\text{-}1}$	t_n x_n	**No. of Differences > 0**	**No. of Differences < 0**
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}		
Total							(S^+)	(S^-)

Q-3.3.3. Example of a Mann-Kendall Trend Test for Small Sample Sizes (n < 10). Evaluate the linear trend of benzene taken from quarterly groundwater samples at well MW01 in Site A from 2000–2001.

Q-3.3.3.1. Benzene has been detected during all of these sampling events, so no proxy concentrations were derived. At the 90% level of confidence ($\alpha = 0.10$), test:

H_0: No trend; H_A: Downward trend.

Q-3.3.3.2. Figure Q-4 is a plot of the concentrations over time. It does appear that a downward trend is present. This test, though, will identify if a statistically significant trend is present (Table Q-4).

Q-3.3.3.3. The Mann-Kendall test statistic, S = 5 – 16 = –11.

Q-3.3.3.4. Using Table B-10 of Appendix B, the p value for n = 7 and $|S|$ = 11 is p = 0.068.

Q-3.3.3.5. As S < 0 and $p < \alpha = 0.10$, we reject H_0 and conclude there is significant evidence of a downward trend.

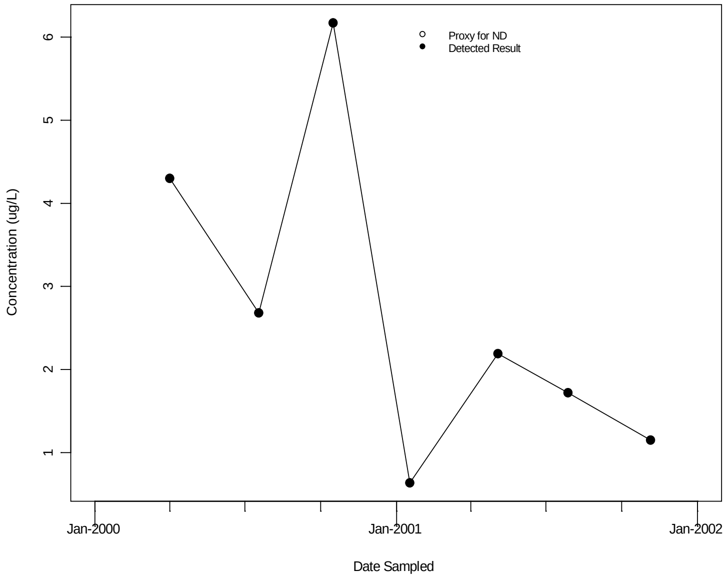


Figure Q-4. Trend for Benzene in Groundwater (small sample size).

Q-3.3.4. <u>Directions for a Normal Approximation to the Mann-Kendall Test Procedure</u>. List the data in the order collected over time. Assign a proxy value to values reported as below the DL. Note that this proxy value should be lower than any measured value. Construct a Data Matrix similar to the top half of the data table below (Table Q-5).

Q-3.3.4.1. Compute the sign of all possible differences as shown in the bottom portion of Table Q-5.

Q-3.3.4.2. Compute the Mann-Kendall statistic, *S*, as shown in Paragraph Q-3.3.2. *S* is the number of positive signs minus the number of negative signs in the triangular table: $S = S^{+} - S^{-}$.

Q-3.3.4.3. If there are no ties, calculate the variance of *S*:

$$V(S) = \frac{n(n-1)(2n+5)}{18}.$$

Q-3.3.4.4. If ties occur, let *g* represent the number of tied groups and w_j represent the number of data points in the j^{th} tied group. For ties, the variance of *S* is:

$$V(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{j=1}^{g} w_j(w_j - 1)(2w_j + 5) \right].$$

Table Q-4.
"Upper Triangular" Data for Basic Mann-Kendall Trend Test with a Single Measurement at Each Time Point—Data Table

Time	7/00	10/00	1/01	5/01	7/01	11/01	**No. of Differences > 0**	**No. of Differences < 0**
$\boldsymbol{x_i}$	2.68	6.17	0.64	2.19	1.72	1.15		
$x_1 = 4.3$	–1.62	1.87	–3.66	–2.11	–2.58	–3.15	1	5
$x_2 = 2.68$		3.49	–2.04	–0.49	–0.96	–1.53	1	4
$x_3 = 6.17$			–5.53	–3.98	–4.45	–5.02	0	4
$x_4 = 0.64$				1.55	1.08	0.51	3	0
$x_5 = 2.19$					–0.47	–1.04	0	2
$x_6 = 1.72$						–0.57	0	1
Total							***5***	***16***

. Q-3.3.4.5. Calculate the following statistic:

$$z = \begin{cases} \dfrac{S-1}{\sqrt{V(S)}}, & S > 0 \\ 0, & S = 0 \\ \dfrac{S+1}{\sqrt{V(S)}}, & S < 0 \end{cases}$$

Q-3.3.4.6. Note that tied values do not affect the calculation of *S* but affect only *V*(*S*) and the calculation of *z* using the large sample approximation.

Q-3.3.4.7. Use Table B-15 of Appendix B to find the critical value $Z_{1-\alpha}$ (if testing H_0: No trend against H_A: Upward trend) or the critical value $-Z_{1-\alpha}$ (if testing H_0, no trend against H_A: downward trend) such that (1 – α)100% of the normal distribution lies to the left of $Z_{1-\alpha}$.

Q-3.3.4.7.1. For testing H_0, no trend against H_A: upward trend, reject H_0 if $z > Z_{1-\alpha}$.

Q-3.3.4.7.2. For testing H_0, no trend against H_A: downward trend, reject H_0 if $z < -Z_{1-\alpha}$.

Table Q-5.
Data for Example Q-3.3.5

Jun-98	Apr-98	Jul-98	Oct-98	Apr-99	Jul-99	Oct-99	Apr-00	Jul-00	Oct-00	Jan-01	May-01	Jul-01	Nov-01	**Time: earliest to latest**	
x_i	3.79	3.42	5.47	0.81	1.78	7.56	4.3	2.68	6.17	0.64	2.19	1.78	1.15	Benzene concentrations	
														#of + Diff.	#of – Diff.
12.2	–8.41	–8.78	–6.73	–11.4	–10.4	–4.6	–7.9	–9.52	–6.0	–11.6	–10.0	–10.4	–11.1	0	13
3.79		–0.37	1.68	–2.98	–2.01	3.77	0.51	–1.11	2.38	–3.15	–1.6	–2.01	–2.64	4	8
3.42			2.05	–2.61	–1.64	4.14	0.88	–0.74	2.75	–2.78	–1.23	–1.64	–2.27	4	7
5.47				–4.66	–3.69	2.09	–1.17	–2.79	0.7	–4.83	–3.28	–3.69	–4.32	2	8
0.81					0.97	6.75	3.49	1.87	5.36	–0.17	1.38	0.97	0.34	8	1
1.78						5.78	2.52	0.90	4.39	–1.14	0.41	0.00	–0.63	5	2
7.56							–3.26	–4.88	–1.39	–6.92	–5.37	–5.78	–6.41	0	7
4.3								–1.62	1.87	–3.66	–2.11	–2.52	–3.15	1	5
2.68									3.49	–2.04	–0.49	–0.90	–1.53	1	4
6.17										–5.53	–3.98	–4.39	–5.02	0	4
0.64											1.55	1.14	0.51	3	0
2.19												–0.41	–1.04	0	2
1.78													–0.63	0	1
1.15												***Total***		***28***	***62***

Q-3.3.5. Example of The Mann-Kendall Procedure Using Normal Approximation for Larger Samples. Consider evaluating whether or not there is a significant trend for benzene using a set of samples taken from quarterly groundwater samples at well MW01 in Site A from 1998–2001. Benzene has been detected during all of these sampling events, so no proxy concentrations were derived.

Q-3.3.5.1. Test H_0, no trend against H_A: downward trend based on a 90% level of confidence ($\alpha = 0.10$).

Q-3.3.5.2. Figure Q-5 is a plot of the concentrations over time. It does appear that a downward trend is present (Table Q-5).

Q-3.3.5.3. The Mann-Kendall statistic, $S = 28 - 62 = -34$.

Q-3.3.5.4. Since there are two observations with a value of 1.78, there are $g = 1$ tied groups and $w_1 = 2$. The calculated variance of S is

$$V(S) = \frac{n(n-1)(2n+5) - \sum_{j=1}^{g} w_j\left(w_j - 1\right)\left(2w_j + 5\right)}{18}$$

$$= \frac{14(13)(33) - 2(1)(9)}{18} = 332.7\ .$$

Q-3.3.5.5. Because $S < 0$, the approximate z test statistic is

$$z = \frac{S+1}{\sqrt{V(S)}} = \frac{-34+1}{\sqrt{332.7}} = -1.809.$$

Q-3.3.5.6. Using Table B-15 of Appendix B, find the critical value $-Z_{0.90} = -1.28$.

–1.809 < –1.28, so we can reject H_0.

Q-3.3.5.7. That means there is significant evidence of a downward trend.

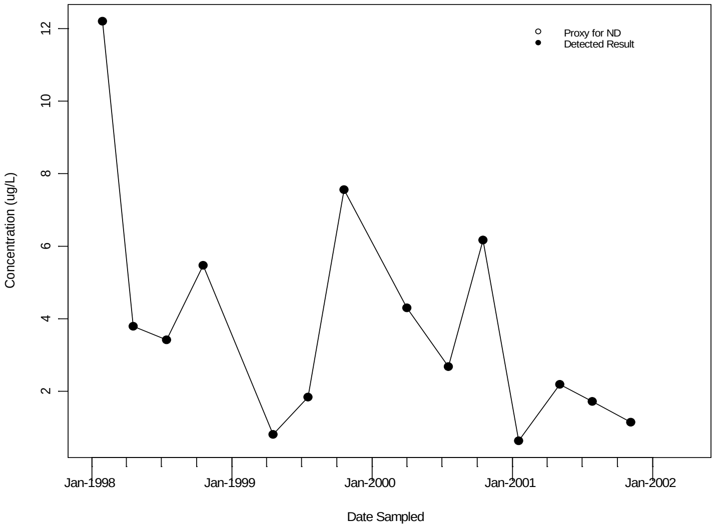


Figure Q-5. Trend for Benzene in Groundwater (large sample size).

Q-3.3.6. Multiple Observations. Often, more than one sample is collected for each time period. There are two ways to deal with such multiple observations. 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 to the summary statistic. The summary statistic would be used instead of the individual data points in the triangular table. The steps given for the Mann-Kendall for small sample sizes or larger samples could then be applied to the summary statistics.

Q-3.3.6.1. An alternative approach is to consider all of the multiple observations within a given time period as being essentially equal (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 is changed to:

$$V(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{j=1}^{g} w_j(w_j-1)(2w_j+5) - \sum_{k=1}^{h} u_k(u_k-1)(2u_k+5)\right]$$
$$+ \frac{\sum_{j=1}^{g} w_j(w_j-1)(w_j-2)\sum_{k=1}^{h} u_k(u_k-1)(u_k-2)}{9n(n-1)(n-2)} + \frac{\sum_{j=1}^{g} w_j(w_j-1)\sum_{k=1}^{h} u_k(u_k-1)}{2n(n-1)}$$

where *g* represents the number of tied groups (i.e., number of groups that have tied observations), w_j represents the number of data points in the tied j^{th} group, *h* is the number of time periods that contain multiple data, and u_k is the sample size in the k^{th} time period where k = 1, 2, …, *h*. For example, let four *X* measurements be made for the first time period, three for the second, two for the third, and one for each of the subsequent time periods. The value of *h* will be 3 for the three time periods with multiple measurements, and the value of u_k will be 4, 3, and 2 for k = 1, 2, and 3 respectively. The values of *g* and w_j will depend on actual *X* measurements. For the special case of ties and multiple measurements for a time period, the reader is referred to Gilbert (1987).

Q-3.3.6.2. 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 (i.e., the median) for each period and then apply either the Mann-Kendall for small sample sizes or larger samples to the summary statistics.

Q-3.3.6.3. The preceding methods involve a single sampling location (station). However, environmental data often consist of sets of data collected at several sampling locations (e.g., groundwater monitoring wells). 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 may be desirable to simultaneously evaluate all of the sampling locations for the presence of a common characteristic or "regional trend." However, there must be consistency in behavioral characteristics across sites over time 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 of the measurements from all stations in which a different symbol is used to represent each station. Paragraph Q-3.3.7 illustrates such data sets.

Q-3.3.6.4. If the stations exhibit approximately steady trends in the same direction (upward or downward), with comparable slopes, a single summary statement across stations is valid, implying that two relevant sets of hypotheses should be investigated.

Q-3.3.6.4.1. Comparability of Stations.

H_0: The trends at all K stations are homogeneous.

H_A: At least two stations exhibit different dynamics.

Q-3.3.6.4.2. Testing for Overall Monotonic Trend.

H_0^* : Contaminant levels do not change over time.

H_A^*: There is an increasing or decreasing trend consistently exhibited across all stations.

Q-3.3.6.5. Therefore, the analyst must first test for homogeneity of stations and then, if homogeneity is confirmed, test for an overall monotonic trend.

Q-3.3.6.6. Ideally, the stations 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. The EPA recommends that an equal number of observations (a balanced design) be required for fewer than three time periods. For four or more time periods, up to one missing value per sampling location may be tolerated.

Q-3.3.6.7. 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. Directions for this condition are presented in Paragraph Q-3.3.8, followed by an example in Paragraph Q-3.3.9.

Q-3.3.6.8. Gilbert (1987) states: “The validity of these chi-squared tests depends on each of the z_k values having a standard normal distribution. [T]his implies that the number of data (over time) for each station should exceed 10. Also, the validity of the tests requires that the z_k values be independent, meaning data from different stations must be uncorrelated.”

Q-3.3.6.9. If multiple measurements are taken at some time and station, the previous approaches are still applicable. However, the variance of the statistic S_k must be calculated using the equation for calculating $V(S)$ based on multiple observations within a given time period. Note that S_k is computed for each station, so n, w_j, g, h, and u_k are all station-specific.

Q-3.3.7. Illustration of Data Taken from Multiple Stations and Multiple Times. Let $i = 1, 2,..., n$ represent time, let $k = 1, 2,..., K$ represent sampling locations or stations, and $x_{i,k}$ represent the measurement at time i for location k. These data can be summarized in matrix form, as shown below:

		Station			
		1	2	…	K
Time	1	$x_{1,1}$	$x_{2,1}$	…	$x_{K,1}$
	2	$x_{1,2}$	$x_{2,2}$	…	$x_{K,2}$
	.	.	.	…	.
	.	.	.	…	.
	.	.	.	…	.
	n	$x_{1,n}$	$x_{2,n}$	…	$x_{K,n}$
		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

$V(S_k)$ = variance for S statistic for station k

z_k = $S_k\big/\sqrt{V(S_k)}$.

Q-3.3.8. Directions for the Mann-Kendall Statistic Used to Test a Monotonic Trend. Let i = 1, 2,..., n represent time, k = 1, 2,..., K represent sampling locations or stations, and $x_{i,k}$ 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 an overall trend.

Q-3.3.8.1. Calculate the Mann-Kendall statistic S_k and its variance $V(S_k)$ for each of the K stations using the methods for larger sample sizes.

Q-3.3.8.2 For each of the K stations, calculate

$$z_k = S_k\big/\sqrt{V(S_k)}\ .$$

Q-3.3.8.3 Calculate the average

$$\bar{z} = \sum_{k=1}^{K} z_k\big/K\ .$$

Q-3.3.8.4. Calculate the homogeneity chi-square statistic

$$\chi_h^2 = \sum_{k=1}^{K} z_k^2 - (K\,\bar{z}^2)\,.$$

Q-3.3.8.5. Using a chi-squared table, find the critical value, $\chi^2_{1-\alpha,\nu}$ the (1 – α)100th percentile of the chi-squared distribution with $\nu = K - 1$ degrees of freedom.

Q-3.3.8.5.1. If $\chi_h^2 > \chi_{1-\alpha,\nu}^2$, the stations are not homogeneous (have 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 previously. That is, test each of the K wells individually as described in Paragraphs Q-3.3.3 or
Q-3.3.5.

Q-3.3.8.5.2. If $\chi_h^2 \le \chi_{1-\alpha,\nu}^2$, there are comparable dynamics across stations at significance level α. Using a chi-squared table, find the critical value for the chi-squared distribution with 1 degree of freedom at the $\alpha *$ significance level, $\chi_{1-\alpha *,1}^2$.

Q-3.3.8.6. If $K\bar{z}^2 > \chi_{1-\alpha *,1}^2$, then reject H_0^* and conclude that there is a significant (upward or downward) monotonic trend across all stations at significance level $\alpha *$. The signs of the S_k indicate whether increasing or decreasing trends are present.

Q-3.3.8.7. If $K\bar{z}^2 \le \chi_{1-\alpha *,1}^2$, there is not significant evidence at the $\alpha *$ level of a monotonic trend across all stations; that is, the stations appear approximately stable over time.

Q-3.3.9. <u>Example of Comparability of Stations and an Overall Monotonic Trend</u>. The following wells at Site A are to be evaluated to determine if the benzene concentrations show decreasing trends consistently across these wells based on a 95% level of confidence. Data for benzene at these wells are shown in the Table Q-6. The flag “ND” is applied to sample for which benzene was not detected. For non-detected concentrations, proxy values are presented in the table and are set to the sample’s detection limit.

Q-3.3.9.1. For this example, $K = 3$.

Q-3.3.9.2. The average of the z values is

$$\bar{z} = (-1.916 - 1.040 + 2.135)/3 = -0.2737\,.$$

Q-3.3.9.3. The homogeneity chi-square statistic is

$$\chi_h^2 = \sum_{k=1}^{K} z_k^2 - (K\,\bar{z}^2) = \left[(-1.916)^2 + (-1.040)^2 + (2.135)^2\right] - 3(-0.2737)^2 = 9.086\,.$$

Q-3.3.9.4. The critical value is $\chi_{0.95,2}^2 = 5.991$, with $\nu = K - 1 = 2$ degrees of freedom and 95% level of confidence (from Table B-2 of Appendix B).

Q-3.3.9.5. Because $\chi_h^2 \ge \chi_{0.95,2}^2$, the stations are not homogeneous based on a 95% level of confidence, and each should be tested using the technique presented in Paragraph Q-3.3.5 as n > 10.

Table Q-6.
Benzene Data for Example Q-3.3.9

Time	Well (Site A)					
	MW01		MW03		MW05	
1	12.2		0.062	ND	2.17	
2	3.79		1.78		2.75	
3	3.42		0.04	ND	6.91	
4	5.47		2.31		8.64	
5	0.81		7.24		11.0	
6	1.84		1.85		14.1	
7	7.56		0.31		3.45	
8	4.30		2.00		36.7	
9	2.68		0.14		20.2	
10	6.17		0.23		8.34	
11	0.64		0.065	ND	17.0	
12	2.19		0.76		21.8	
13	1.72		0.22		2.01	
14	1.15		0.05	ND	29.1	
S_k	–35		–19		39	
$V(S_k)$	333.7		333.7		333.7	
z_k	–1.916		–1.040		2.135	

Q-3.3.10. Multiple Observations over Extended Time Periods. Temporal data are often collected over extended time periods. Within the time variable, data may exhibit periodic cycles, patterns in the data that repeat over time. For example, temperature and humidity may change with the season or month and affect environmental measurements. For this 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. There are two approaches for testing for trends—the seasonal Kendall test and Sen’s test for trends—if seasonal cycles are anticipated. The seasonal Kendall test may be used for large sample sizes, and Sen’s test for trends may be used for small sample sizes. In either case, the data are analyzed separately by season, and the results are compared among seasons. Both of these estimation techniques are described below. If different seasons manifest similar slopes (rates of change) but different intercepts, the Mann-Kendall technique for multiple sampling locations with multiple observations is applicable, replacing station by season. For example, Figure Q-6 shows a time plot of a series that appears to be decreasing although it is somewhat masked by a seasonal cycle that repeats every four time periods. The data could be analyzed by the Mann-Kendall technique presented in Paragraph Q-3.3.8 if they are broken out by season (e.g., data points 1, 5, 9, 13, and 17 would constitute one season series).

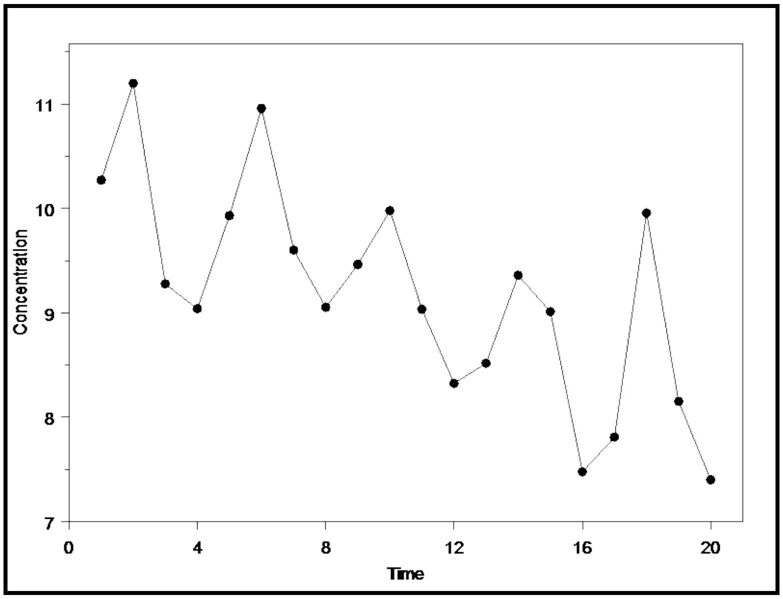


Figure Q-6. Time Plot of Seasonal Series with Decreasing Trend.

Q-3.3.10.1. For data with seasonality, the seasonal Kendall test, 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.

Q-3.3.10.2. For data at a single site, collected at multiple seasons within multiple years, the techniques for multiple sampling locations with multiple observations can be used to test for homogeneity of time trends across seasons. The methodology follows the explanation below of Sen's slope estimator exactly, except "station" is replaced by "season" and the inferences refer to seasons.

Q-3.3.10.3. If a linear trend is observed when some variable of interest is plotted against time, based on a visual inspection or the results of a statistical test for a trend, the magnitude of the slope of the line is a measure of the "strength" of the trend and the sign of the slope provides the direction of the trend. The true slope (change per unit time) may be estimated using a parametric or non-parametric method. Linear regression analysis is a parametric method for estimating a slope. Sen's slope estimator is a non-parametric method for estimating the slope of a line.

Q-3.3.10.4. This approach involves computing slopes for all pairs of ordinal time points and 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.

Q-3.3.10.5. Directions are presented in Paragraph Q-3.3.11, followed by an example in Paragraph Q-3.3.12.

Q-3.3.11. Directions for a Sen's Slope Estimator. Assume that there are *n* time points (or *n* periods of time), and let x_i denote the data value for the i^{th} time point. If there are no missing data, there will be $N' = n(n-1)/2$ possible pairs of time points (*i*, *j*), in which $i > j$ (i.e., x_i was taken at a time after the measurement x_j).

Q-3.3.11.1. For non-detected results, the detection limit may be used as the data value (Gibbons, 1994) or one-half the detection limit may be used as the data value (Gilbert, 1987). Note that this proxy value should be lower than any measured value.

Q-3.3.11.2. Define the slope for each pair, called a pairwise slope, as

$$b_{ij} = \frac{(x_i - x_j)}{(i-j)}.$$

Q-3.3.11.3. Sen's slope estimator is the median of the $n(n-1)/2$ pairwise slopes.

Q-3.3.12. Example of a Sen's Slope Estimator. The Sen's slope estimate is calculated to evaluate the linear trend for benzene in Paragraph Q-3.3.3 (seven groundwater samples collected quarterly from 2000–2001 from well MW01 at Site A). Because benzene was detected for all the sampling events, proxy concentrations were not derived.

Q-3.3.12.1. There are 7(6)/2 = 21 possible pairs of time points (*i*, *j*) in which $i > j$. The slope for each pair will be estimated and displayed in a data matrix similar to the one presented in Paragraph Q-3.3.3, except each cell in the matrix represents the pairwise slope

$$b_{ij} = \frac{(x_i - x_j)}{(i-j)}.$$

Q-3.3.12.2. If there is no underlying trend, then a given x_i is just as likely to be above another x_j as it is to be below. If there is no underlying trend, there would be an approximately equal number of positive and negative slopes and Sen's slope would be near zero.

Q-3.3.12.3. If the data exhibit cyclic trends, the 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 and the length of a cycle is one year, 12 separate sets of slopes would be determined (one for each month of the year using all of the data for that particular month); 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, which is rarely calculated by hand owing to the number of calculations required.

Table Q-7.
Pairwise Slopes Data Table

Original							
Time	*t_1=4/00*	*t_2=7/00*	*t_3=10/00*	*t_4=1/001*	*t_5=5/01*	*t_6=7/01*	*t_7=11/01*
Measure	*x_1=4.3*	*x_2=2.68*	*x_3=6.17*	*x_4=0.64*	*x_5=2.19*	*x_6=1.72*	*x_7=1.15*
x_1=4.3		–1.62	0.935	–1.22	–0.528	–0.516	–0.525
x_2=2.68			3.49	–1.02	–0.163	–0.24	–0.306
x_3=6.17				–5.53	–1.99	–1.483	–1.255
x_4=0.64					1.55	0.54	0.17
x_5=2.19						–0.47	–0.52
x_6=1.72							–0.57
x_7=1.15							

Ordered pair-	–5.53	–1.99	–1.62	–1.483	–1.255	–1.22	–1.02
wise slopes	–0.57	–0.528	–0.525	–0.52	–0.516	–0.47	–0.306
(smallest to	–0.24	–0.163	0.17	0.54	0.935	1.55	3.49
largest):							

Q-3.3.12.4. The median of these 21 pairwise slopes is –0.52, the 11th ordered result when the results are sorted from smallest to largest.

Q-3.3.13. Testing a Trend Using Confidence Limits for Sen's Slope Estimator. Gilbert (1987) presents a simple method, based on the normal distribution, to estimate the (1 – α)100% confidence interval about the true slope. This "large sample" estimate is appropriate for data sets with at least 10 samples. Directions for estimating such confidence intervals are presented below. Aside from estimating the confidence limits for the slope associated with a trend that has been previously identified (e.g., using Mann-Kendall's test), this approach can be used to determine if a trend is presented. If the confidence interval for the slope contains zero, there is no evidence of an underlying trend. However, if the confidence interval does not contain zero, there is evidence to suggest a trend. Directions are presented in Paragraph Q-3.3.14, followed by an example in Paragraph Q-3.3.15.

Q-3.3.14. Directions for Creating Confidence Limits for Sen's Slope Estimator. Compute $N'=n(n-1)/2$ if there is just one result in each time period, and $N'=$ the number of possible data pair combinations among the time periods (and results from the time period cannot be considered data pairs) if there is more than one result in each time period.

Q-3.3.14.1. Based on the desired two-sided confidence level (1 – α)100%, find $Z_{1-\alpha/2}$.

Q-3.3.14.2. Compute the variance of S as

$$V(S)=\frac{1}{18}\left[n(n-1)(2n+5)-\sum_{j=1}^{g}w_j(w_j-1)(2w_j+5)\right]$$

when one observation per time period is available (g represents the number of tied groups and w_j represent the number of data points in the j^{th} group) or

$$V(S)=\frac{1}{18}\left[n(n-1)(2n+5)-\sum_{j=1}^{g}w_j(w_j-1)(2w_j+5)-\sum_{k=1}^{h}u_k(u_k-1)(2u_k+5)\right]$$
$$+\frac{\sum_{j=1}^{g}w_j(w_j-1)(w_j-2)\sum_{k=1}^{h}u_k(u_k-1)(u_k-2)}{9n(n-1)(n-2)}+\frac{\sum_{j=1}^{g}w_j(w_j-1)\sum_{k=1}^{h}u_k(u_k-1)}{2n(n-1)}$$

when multiple observations per time period are available (g represents the number of tied groups, w_j represents the number of data points in the j^{th} group, h is the number of time periods containing multiple data, and u_k is the sample size in the k^{th} time period).

Q-3.3.14.3. Compute $C_\alpha = Z_{1-\alpha/2}\sqrt{V(S)}$.

Q-3.3.14.4. Compute $M_1 = (N'-C_\alpha)/2$ and $M_2 = (N'+C_\alpha)/2$.

Q-3.3.14.5 The lower and upper limits of the confidence interval are the M_1^{th} largest and $(M_2 + 1)^{th}$ largest of the N' ordered slope estimates (from lowest to highest), respectively. If M_1 and $M_2 + 1$ are not whole numbers, use linear interpolation (Gilbert, 1987).

Q-3.3.15. Example of Confidence Limits for Sen's Slope Estimator. Consider estimating a two-sided 95% confidence interval for Sen's slope estimated in Paragraph Q-3.3.12, where:

$n = 7$, $S = -0.52$ and $N' = n(n-1)/2 = 21$.

Q-3.3.15.1. For $\alpha = 0.05$, $Z_{1-\alpha/2} = Z_{0.975} = 1.96$.

Q-3.3.15.2. The following are calculated:

$$V(S)=\frac{1}{18}\left[n(n-1)(2n+5)-\sum_{j=1}^{g}w_j(w_j-1)(2w_j+5)\right]=\frac{1}{18}[7(6)(19)-0]=44.33$$

$$C_\alpha = Z_{1-\alpha/2}\sqrt{V(S)} = 1.96\sqrt{44.33} = 13.05$$

$$M_1 = (N' - C_\alpha)/2 = (21 - 13.05)/2 = 3.975$$

$$M_2 = (N' + C_\alpha)/2 = (21 + 13.05)/2 = 17.025.$$

Q-3.3.15.3. From the list of ordered results in Paragraph Q-3.3.15 , the interpolated value between the 3rd and 4th ordered result is –1.486 and the interpolated value between the 18th (17 + 1) and 19th ordered result is 0.550. Therefore, the confidence interval for the slope is (–1.486, 0.550). As this interval contains zero, there is insufficient evidence of an underlying trend (even though the slope of –0.52 suggests a negative trend).

Q-4. Control Charts.

Q-4.1. Introduction. Control charts are a quality control procedure that can be applied to environmental monitoring data, such as data from air or groundwater monitoring systems. Control charts provide a visual means of monitoring constituent concentrations at a given well or location over time, identifying slight or sudden fluctuations over time and detecting deviations from a "state of control." A process is in-control if the observed variation is attributable to small, uncontrollable changes. A process is out-of-control if a relatively large variation is introduced that can be traced to an assignable cause (Kvanli et al., 1996).

Q-4.1.1. Control charts are most frequently used in groundwater monitoring detection programs for intra-well comparisons, in which data are collected for a single well over some period of time. Control charts are useful for areas with no previous contamination because detecting contamination may require a significant change. This is particularly applicable to monitoring down-gradient of waste cells or landfills, because it can highlight whether there has been a release to groundwater. If contamination was historically present, it will take a significant increase in concentrations relative to historical values to show a detection (Gibbons, 1994). Control charts, however, are not constructed for making precise probability statements; they are constructed as a guide for determining when investigative action is needed (Gilbert, 1987). Furthermore, contamination may be present intermittently or may increase in a step function. The absence of an increasing trend does not necessarily support that a release has not occurred.

Q-4.1.2. Control charts are designed for a given constituent and well in which concentrations are plotted against time with horizontal lines called "control limits." Control limits are based on meaningful and sufficient historical data with no outliers and trends over time. As new data become available, those concentrations are also plotted. The EPA recommends, and current RCRA regulations specify, developing control limits with data consisting of at least eight independent samples over a 1-year period. As with most statistical applications, more historical data are desirable but, in practical terms, are rarely available.

Q-4.1.3. The assumptions underlying control charts are that when the process is in-control, data are independent and normally distributed with a fixed mean and constant variance. Independence is crucial. Control charts are not robust with respect to the departure from independence (i.e., when data are correlated). To minimize the possibility that samples are dependent, Gibbons (1994) recommends a sampling frequency of no more than one sample per quarter. To identify serial correlation, a sample's serial correlation coefficient can be calculated. (Details are provided in Appendix O.) A correlogram may be plotted to determine if serial correlation is large enough to create problems. (Details are provided in Appendix J.) A quick method for determining if serial correlation is large is to compare the autocorrelation coefficients to

$$\pm 2\Big/\sqrt{n}$$

where *n* is the number of time periods when data were collected. Autocorrelation coefficients that exceed either of these values require further investigation.

Q-4.1.4. The assumption of normality is not nearly as crucial, but the data's distribution should still be investigated. To achieve normality, data transformations (such as natural-log transformations or square-root transformations) should be applied to sample data, as appropriate. Gilbert (1987) suggests that as long as data are normally distributed and the correlation associated with the data is not too large, control chart methods work well. Gilbert goes on to say that although environmental data are typically non-normal, control charts are still useful for indicating where concentrations are not likely to be from the same distribution as in the past.

Q-4.1.5. Seasonality, a component of the data's variability, should also be considered before control charts are developed. Seasonality can be addressed by removing seasonal effects from the data, if sufficient data are available for at least two seasons of the same type. Removing seasonality was previously discussed in Paragraph Q-2. Gilbert (1987) recommends two other methods to circumvent seasonality issues. If data are available for a number of complete cycles, separate control charts for each season can be prepared. If the data do not span a long duration and the magnitude of the cycles is relatively small, a moving-average control chart may be constructed.

Q-4.1.6. In terms of proxy concentrations appropriate for control charts, Gibbons (1994) suggests that if at least 25% of samples are detections, a proxy concentration based on just the sample-specific method detection limit is adequate for control charts.

Q-4.1.7. Several types of control charts are discussed in this section: Shewart control charts, CUSUM control charts, and Shewart-CUSUM control charts. The advantage to Shewart control charts is that they are immediately sensitive to large changes. The advantage to CUSUM control charts is that they are sensitive to small and gradual changes. Shewart-CUSUM control charts are a combination of the other two. As such, their benefit is that they can detect both sudden and gradual changes in concentrations.

Q-4.2. Shewart Control Charts.

Q-4.2.1. Introduction. Shewart control charts, which are the oldest and simplest charts (Gibbons, 1994), are sensitive to sudden changes and focus on the current monitoring value. Current data (not historical data) are first plotted against time. Control limits are subsequently placed on the same plot as horizontal lines. The control limits are calculated using historical data from a period of time when the system under study was stable. New data that fall outside of the control limits indicate that current conditions have changed from the historical ones used to establish the control limits. Although lower control limits are used in other fields, only the upper control limit is typically established for environmental data, as the objective is to identify dramatically increasing concentrations. An upper control limit can be developed from historical data using the equation $\mu + Z\sigma$, where μ is the population mean, σ is the population standard deviation, and Z is an upper percentage point of the normal distribution. For this case, Z is typically equal to 3, which corresponds to a confidence level of $1-\alpha = 0.9987$ for a single new comparison.

Q-4.2.1.1. However in most cases, long-run historical data are unavailable and a sample estimate of the mean ($\bar{x}$) and standard deviation (*s*) must be used. In this case, the equation for the upper control limit is $\bar{x} + Zs$. When using the sample estimates to calculate an upper control limit with as few as eight historical samples, however, the control limit only provides an overall 95% confidence for five new comparisons and the overall confidence decreases as the number of future observations increases (Gibbons, 1994). As such, EPA 530-SW-89-026 recommends setting control limits to $\bar{x} + 4.5s$ for routine groundwater monitoring situations. "Overall confidence levels for this control limit are 95% with *n* = 8 and 35 future comparisons; however, verification resampling further reduces false positive rates to acceptable levels for most monitoring programs" (Gibbons, 1994), avoiding the problem of multiple comparisons discussed in Appendices M and N. It should be noted that 4.5 is a generic value recommended by the EPA to be protective in most monitoring situations. Gibbons, 1994 warns "[t]he reader should note that unlike prediction limits which provide a fixed confidence level (e.g. 95%) for a given number of future comparisons, control charts do not provide explicit confidence levels, and they do not adjust for the number of future comparisons." See Appendix K for information on developing prediction limits to cover a specific number of future observations and tolerance limits to cover an indefinite number of future observations.

Q-4.2.1.2. If more than eight historical samples are available, it is reasonable to use only the most recent eight. Once a control limit is developed, the current monitoring value is compared to the control limit. If the value exceeds the control limit, the groundwater system should be investigated for causes associated with the increase in concentration. Directions for preparing a Shewart control chart are given in Paragraph Q-4.2.2, followed by an example in Paragraph Q-4.2.3.

Q-4.2.2. Directions for Preparing a Shewart Control Chart.

Q-4.2.2.1. Verify the following assumptions:

Q-4.2.2.1.1. For each sampling location (e.g., a well for groundwater monitoring), data are available from at least eight independent samples from previous sampling events to estimate the mean and standard deviation.

Q-4.2.2.1.2. Determine if data are correlated.

Q-4.2.2.1.3. Identify if data or transformed data are normally distributed.

Q-4.2.2.1.4. Check if seasonality is affecting data, and, if so, remove the seasonality.

Q-4.2.2.2. At a given location or well, take independent samples over *n* historical sampling events ($n \geq 8$).

Q-4.2.2.3. Calculate the mean ($\bar{x}$) and standard deviation (*s*) of the *n* samples.

Q-4.2.2.4. Calculate an upper control limit by the equation $\bar{x} + Zs$, where *Z* is set to 4.5 for routine groundwater monitoring programs. Note that setting $Z = 4.5$ ensures a 95% overall confidence level when $n = 8$ and 35 future comparisons are made to this upper control limit (Gibbons, 1994).

Q-4.2.2.5. Plot the current concentrations with respect to time and superimpose the upper control limit.

Q-4.2.2.6. Identify if the system is in-control or out-of-control by identifying if concentrations are below the upper control limit or above the upper control limit, respectively.

Q-4.2.2.7. Investigate any situation in which a concentration is above the upper control limit.

Q-4.2.3. Example of a Shewart Control Chart. Benzene is measured from quarterly groundwater samples at well MW01 in Site A from 1998–2000 to develop a control chart to compare to the 2001 sampling results (Table Q-8).

Q-4.2.3.1. Verifying assumptions are as follows.

Q-4.2.3.1.1. $n = 10$. Samples were taken with at least a 3-month interval; therefore, the samples should be independent.

Q-4.2.3.1.2. This set of data is the same as that used to calculate the serial correlation for the example in Paragraph O-2.6.2. From that example, the following summary statistics were estimated: $\bar{x} = 4.824$ and $s_x = 3.284$, and the serial correlation coefficient = –0.2527. The correlogram for these data is shown in Figure Q-7.

Q-4.2.3.2. Serial correlation does not appear to be a problem, even though the default at k = 0 (where k is the autocorrelation coefficient) is greater than the $\pm 2/\sqrt{n}$ bounds (± 0.632).

Table Q-8a.
Historical Data for Upper Control Limit in Example Q-4.2.3

Time	Jan-98	Apr-98	Jul-98	Oct-98	Apr-99	Jul-99	Oct-99	Apr-00	Jul-00	Oct-00
Time Period	1	2	3	4	5	6	7	8	9	10
Conc. (µg/L)	12.2	3.79	3.42	5.47	0.81	1.84	7.56	4.32	0.68	6.17

Table Q-8b.
Current Data to Use to Compare to Control Limit in Example Q-4.2.3

Time	Jan-01	May-01	Jul-01	Nov-01
Conc. (µg/L)	0.64	2.19	1.72	1.15

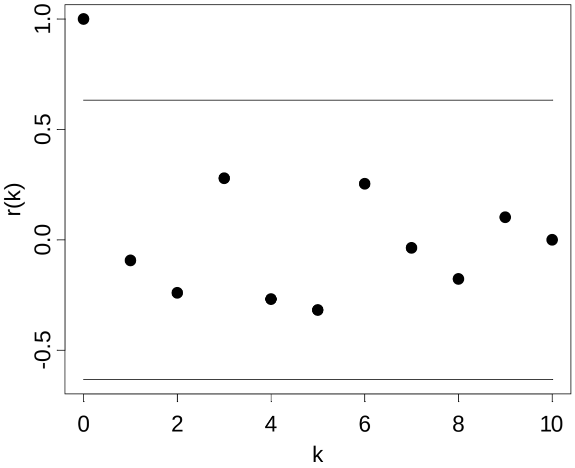


Figure Q-7. The Correlogram.

Q-4.2.3.2.1. To test the assumption of normality, the Shapiro-Wilk test was performed with the data based on a 95% level of confidence. Results of this test provide evidence to suggest that the data follow a normal distribution because the p value is 0.3363, which is greater than the significance level of $\alpha = 0.05$ (there is not enough evidence to reject the null hypothesis of normality).

Q-4.2.3.2.2. There are not enough results to adequately identify seasonal trends and no obvious trend is visible in the previous time plot. For this example, we will assume that the data are not affected by seasonality.

Q-4.2.3.3. There are not enough results to adequately identify seasonal trends and no obvious trend is visible in the previous time plot. For this example, we will assume that the data are not affected by seasonality. Calculate the control limit as follows.

Q-4.2.3.4. The upper control limit $= \overline{x} + Z\,s = 4.824 + (4.5 \times 3.284) = 19.06$. None of the samples taken in 2001 exceeds this upper control limit, as shown in Figure Q-8.

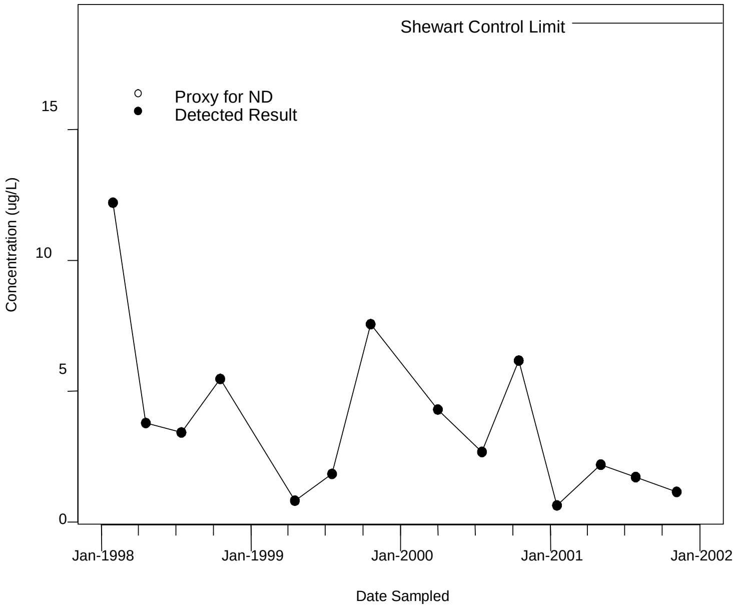


Figure Q-8. Historical Data (1998–2000) and 2001 Data with Shewart Control limit for benzene (SW8260B) in Groundwater at Site A, MW-01.

Q-4.3. CUSUM Control Charts. CUSUM control charts are more sensitive than Shewart control charts to small and gradual changes. They incorporate current and historical information by calculating a cumulative sum, *S*, for the *i*th sample. Directions for preparing a CUSUM control chart are provided in Paragraph Q-4.3.1, followed by an example in Paragraph Q-4.3.2. See Gibbons (1994) for more information.

Q-4.3.1. Directions for a CUSUM Control Chart. Verify that the assumptions required for CUSUM charts are met.

Q-4.3.1.1. Assumptions.

Q-4.3.1.1.1. At least eight independent samples (from previous sampling events) were collected for each sampling location (groundwater monitoring well) to estimate the mean, $\bar{x}$, and sample standard deviation, *s*.

Q-4.3.1.1.2. The data cannot be correlated; determine if the data are correlated.

Q-4.3.1.1.3. The data must be normal; determine whether the data or transformed data are normally distributed.

Q-4.3.1.1.4. Determine whether seasonality is affecting the data; if so, remove the seasonality.

Q-4.3.1.2. Calculate the mean and standard deviation for the historical data results.

Q-4.3.1.3. Choose an appropriate value for *k* (one-half the size of a difference worth detecting). The EPA recommends setting *k* = 1, which means that a difference of two units of standard deviation is meaningful.

Q-4.3.1.4. At a given location or well, determine the cumulative sum for each independent sample. Define

$$S_i = \begin{cases} 0, & i = 0 \\ \max[0,\ z_i - k + S_{i-1}], & i = 1, 2, \ldots, n \end{cases}$$

where $z_i = \dfrac{x_i - \bar{x}}{s}$. The function max[*a*, *b*] means to use the value *a* or *b*, whichever is higher.

Q-4.3.1.5. Choose the appropriate control limit, *h*. EPA recommends setting *h* = 5. The value of 5 is based on simulations and recommendations contained in Lucas (1982), Hockman and Lucas (1987), and EPA 600/4-88-/040. Essentially, *h* is the upper control limit. (One way to determine whether S_i exceeds five is to plot *S* versus *i* for the data.)

Q-4.3.1.6. Identify if the system is in-control or out-of-control by identifying whether each S_i is less than *h* (in-control), or greater than *h* (out-of-control).

Q-4.3.1.7. Investigate any situation in which a concentration is out-of-control. Ideally, additional samples would determine if the out-of-control condition is real and persistent.

Q-4.3.1.8. EPA 530-SW-89-026 recommends detecting a difference of two standard deviations, or *k* = 1. CUSUM control charts are developed by plotting each S_i against the iteration *i*. Each S_i is compared to an appropriate control limit, *h*. EPA guidance recommends $h = 5$. If any S_i value exceeds *h*, the groundwater system should be investigated for causes associated with the increase in concentration.

Q-4.3.2. Preparing a CUSUM Control Chart. Consider evaluating the same data used in the example for developing Shewart Control Charts. Benzene concentrations taken from quarterly groundwater samples at well MW01 in Site A from 1998–2000 will be used as a basis for comparison to the 2001 sampling results.

Q-4.3.2.1. The assumptions for developing CUSUM control charts are the same as developing Shewart control charts. As explained in Paragraph Q-4.2.3, all of these assumptions have been met.

Q-4.3.2.2. Set $k = 1$, $S_0 = 0$, and $h = 5$.

Q-4.3.2.3. For each of the current results in 2001, S_i is calculated as

$$S_i = \max[0,\ z_i - k + S_{i-1}]$$

where

$$z_i = \frac{x_i - \mu}{\sigma}$$

μ is estimated by $\bar{x} = 4.824$, and σ is estimated by $s_x = 3.284$. (Specify what data are being used to calculate the mean and standard deviation.) Each S_i value is then compared to $h = 5$; cases in which $S_i \geq h$ are defined as samples out-of-control. (Note**:** Both the mean and standard deviation come from 10 historical samples.)

Q-4.3.2.4. Results are presented in Table Q-8 and show that none of the current results are out-of-control.

Q-4.3.2.5. As an example of these calculations, consider the July 2001 concentration, where $i = 3$:

$$z_3 = \frac{1.72 - 4.824}{3.284} = -0.945$$

$$S_3 = \max[0,\ (-0.945 - 1 + 0)] = \max[0, -1.945] = 0\,.$$

Q-4.3.2.6. Because $S_3 = 0 < h = 5$, the sample is in-control. In this example, there are no out-of-control events because $S_i < 5$ for all i.

Q-4.4. Combined Shewart-CUSUM Control Charts. Combined Shewart-CUSUM control charts can be used to detect sudden and gradual changes in concentrations. These control charts combine the benefits of the Shewart and CUSUM charts, as illustrated in Paragraph Q-4.4.1.

Q-4.4.1. Consider evaluating the same data used in the example for developing the Shewart and CUSUM control charts. Benzene concentrations taken from quarterly groundwater samples at well MW01 in Site A from 1998–2000 will be used to develop a control chart to compare to the 2001 sampling results.

Q-4.4.2. The assumptions for developing Shewart-CUSUM control charts are the same as for developing Shewart and CUSUM control charts. As explained above, all of these assumptions have been met.

Q-4.4.3. Set $h = 5$, $k = 1$, and use the Shewart chart control limit $\text{SCL} = 4.5$ as recommended by the EPA.

Q-4.4.4. The standardized values for each of the current results are estimated, as shown in Table Q-9. The standardized values, z_i, are developed using the historical average and standard deviation of $\bar{x} = 4.824$ and $s = 3.284$.

Q-4.4.5. Then, each z_i value is compared to $\text{SCL} = 4.5$, and each S_i value is compared to $h = 5$. If $z_i > \text{SCL}$ or $S_i > h$, the result is out-of-control.

Q-4.4.6. Results are presented in Table Q-9 and indicate that none of the current results are out-of-control.

Q-4.4.7. As an example of these calculations, consider the July 2001 concentration, where $i = 3$:

$$z_3 = \frac{1.72 - 4.824}{3.284} = -0.945 \ .$$

Q-4.4.8. $S_3 = \max[0, (-0.945 - 1 + 0)] = \max[0, -1.945] = 0$.

Q-4.4.9. As $z_3 = -0.945 < SCL = 4.5$ and $S_3 = 0 < h = 5$, this sample is in-control.

Q-4.4.10. A plot of the standardized results (z_i) versus the time interval (i) can be designed to illustrate this information, as shown in Figure Q-9.

Table Q-9.
Results for Combined Shewart-CUMSUM Control Chart

Time	Jan-01	May-01	Jul-01	Nov-01	
Concentration (µg/L)	0.64	2.19	1.72	1.15	
i	1	2	3	4	
z_i	–1.274	–0.802	–0.945	–1.119	Compare z_i to *SCL* = 4.5.
S_i	0	0	0	0	Compare S_i to *h* = 5.
Out-of-control? (i.e., z_i > SCL = 4.5 or, S_i > *h* =5)?	No	No	No	No	

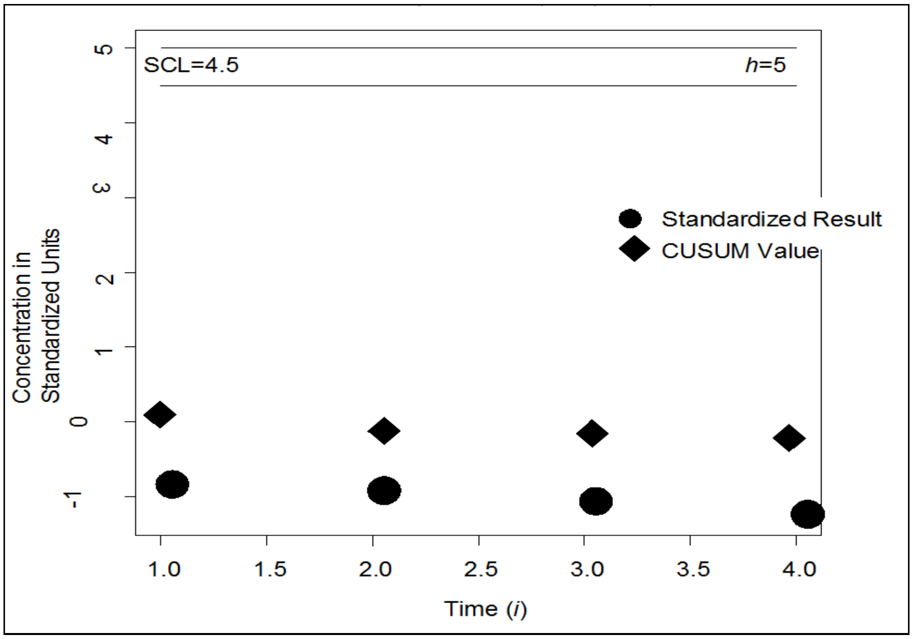


Figure Q-9. Combined Shewart-CUSUM Control Chart (mean = 4.824, standard deviation = 3.284, *k* = 1, *h* = 5, SCL = 4.5).

APPENDIX R

Geostatistics

R-1. Introduction. Geostatistics is a method for analyzing spatially correlated data. It is used to identify spatial patterns and to interpolate values at unsampled locations. Sampling and mapping in the earth sciences are complicated by spatial and temporal patterns. The structure and intensity of such patterns often cannot be reliably predicted with deterministic models of fate and transport or with classical statistical methods applied to sample observations. Geostatistics is a way of interpreting patterns from sample observations taking advantage of spatial correlation. In geosciences, spatial correlation arises when samples taken close to one another are more likely to have similar values than samples taken far apart Clark (1979).

R-1.1. Appendix O explains that covariance is a statistical measure of the association between two variables. If two variables are independent, the covariance is zero. For geostatistical analysis conducted on a regionalized variable, the auto-covariance between nearby samples is considered to be possibly not equal to zero. If the auto-covariance between two measurements taken close to each other is not zero, then the application of classical statistical methods may impart a substantial bias to the estimate.

R-1.2. Classical statistical methods rely on data being independent over distance or time. Hence, in many environmental problems, the use of classical statistics is not entirely accurate, because variables are frequently spatially controlled. Geostatistics recognizes the spatial correlation and provides methods for the following.

R-1.2.1. Calculating predictions (such as the concentration of a metal at a specific location in soil).

R-1.2.2. Quantifying the accuracy of the predictions.

R-1.2.3. Selecting optimal locations to sample given an opportunity to collect more data.

R-1.3. A geostatistician’s main task is to predict a regionalized variable (e.g., hydraulic gradient or metal concentration in soil) from a set of measurements. More detailed treatment of geostatistical methods can be found in Cressie (1993) and Goovaerts (1997).

R-2. Semivariogram. The characteristic tool in geostatistics is the semivariogram to quantify and model the spatial correlation structure. A semivariogram is essentially a plot of the variance of groups of paired sample measurements as a function of the distance between samples. Typically, for the situation in which the variance depends only upon distance (and not direction), all possible sample pairs a fixed distance apart (*h*) are used to calculate a variance for *h*:

$$s^2(h) = \frac{\sum_{i,j}^{N_h}(x_i - x_j)^2}{N_h}$$

where x_i and x_j represent the value (i.e., concentration) at a pair of sample points *i* and *j*; the summation is over all possible pairs of points within a subgroup of the data that are a distance *h* apart (where $i < j$); and N_h denotes the total number of pairs that are *h* units part. For example, *h* is typically defined as increasing along constant intervals ($h = \{1d, 2d, 3d, \ldots\}$, where *d* is a distance interval such as 5 feet.) In practice, a window of allowable distances is used so that many points will be included in each calculation of $s^2(h)$. For example, a group of samples used to calculate s^2 (3 feet) may have inter-point distances that are between 2 feet to 4 feet apart, rather than exactly 3 feet apart. This window is defined using a tolerance δ for *h*, so that all points within $h \pm \delta$ of each other are grouped into the subset from which $s^2(h)$ is calculated. The user chooses this tolerance and other grouping parameters to define how the data will be grouped into subsets to calculate the $s^2(h)$ for each *h*. Different experimental variograms can be calculated for a given data set by varying the grouping parameters used to control the spatial geometry of the data subsets at each distance *h*.

R-2.1. With grouping parameters defined, computer software is used to do the intensive computations involved in calculating the variance $s^2(h)$ for different values of *h*. The quantity $\gamma(h) = (1/2)s^2(h)$ is plotted as a function of increasing distance, i.e., 1*h*, 2*h*, etc., and is referred to as the experimental or empirical semivariogram. Although the *variogram* is, by definition, twice the semivariogram, the terms variogram and semivariogram are often used interchangeably.

R-2.2. After experimental semivariograms are reviewed, a continuous mathematical curve, called a model semivariogram, is then fit to the experimental semivariogram. Examples of model semivariograms are displayed in Figure R-1. The model semivariogram (Figure R-1a) is assumed to characterize the relationship of how variance in neighborhoods increases as the neighborhoods get larger. This relationship must be estimated for each site application. In practice, 20 or more sample locations are necessary to construct a useful empirical semivariogram, and often geological site knowledge and statistical judgment are important considerations in estimating the model semivariogram.

R-2.3. Figures R-1b and R-1c illustrate two model forms that have a sill, or maximum variance. A sill is the upper limit of any semivariogram model that levels off at large distances. In physical terms, the sill is the variance of concentrations at the site that are at a large enough distance from each other to be statistically independent. The distance at which spatial correlation becomes insignificant is called the range. Sample points separated by this distance or more are considered statistically independent and can be analyzed using a classic statistical approach. Another feature of a semivariogram illustrated in Figure R-1c is the nugget. In a model having a nugget, $\gamma(h)$ does not approach zero as h approaches zero but rather a positive value that is generally attributed to such things as measurement error for a single observation or small-scale variability.

R-3. Kriging. The geostatistical interpolation method of kriging uses the concepts and model established in a semivariogram data evaluation to develop both an unbiased estimate of the expected value at any specified location, as well as the uncertainty associated with this estimate. Typically, estimates are derived along a regularly spaced grid. With relatively dense grids, the estimate at each grid point is also the estimate of the mean value within the block centered on the grid point.

R-3.1. The kriging estimate is a weighted mean of the neighboring samples, where each weight reflects the amount of unique (non-redundant) information contained about the location to be estimated that is in a given sample. The assignment of weights to each neighboring sample is based on the model semivariogram, and includes consideration of the inter-point distance between the sample and the location to be estimated, as well as the inter-point distances between this sample and its neighboring samples. Neighboring samples, if close together, are spatially correlated and, therefore, contain redundant (non-independent) information about the location to be estimated. Kriged estimates are more accurate than an unweighted arithmetic mean; that is, they are unbiased (the bias from clustered samples is removed), and they have a lower variance.

R-3.2. Some types of kriging that may be encountered are ordinary kriging, indicator or probability kriging, and block kriging. Ordinary kriging is used to predict the value of some variable at a specific location. In block kriging, the technique allows the prediction of a variable mean within a block or area.

R-3.3. The required assumptions for kriging are that the sample to be estimated lie within the neighborhood for which the model semivariogram has been estimated, that there be adequate empirical evidence (sample data) or scientific support (e.g., source history) for the appropriateness of the model semivariogram, and that the neighborhood be homogeneous, with no distinct trends in the data values. For kriging, a trend is a deterministic gradient that can be modeled (such as an exponential decrease in deposition with distance from a point release). Such trends should be characterized and then subtracted from the regionalized variable being modeled. Kriging can then be run on the residuals to account for local patchiness and clustered sample data. Alternatively a release or plume of contamination can often be divided into strata in which the conditions are approximately homogenous (e.g., geological strata, differing source areas). The blocks of each neighborhood are then kriged using their corresponding semivariogram.

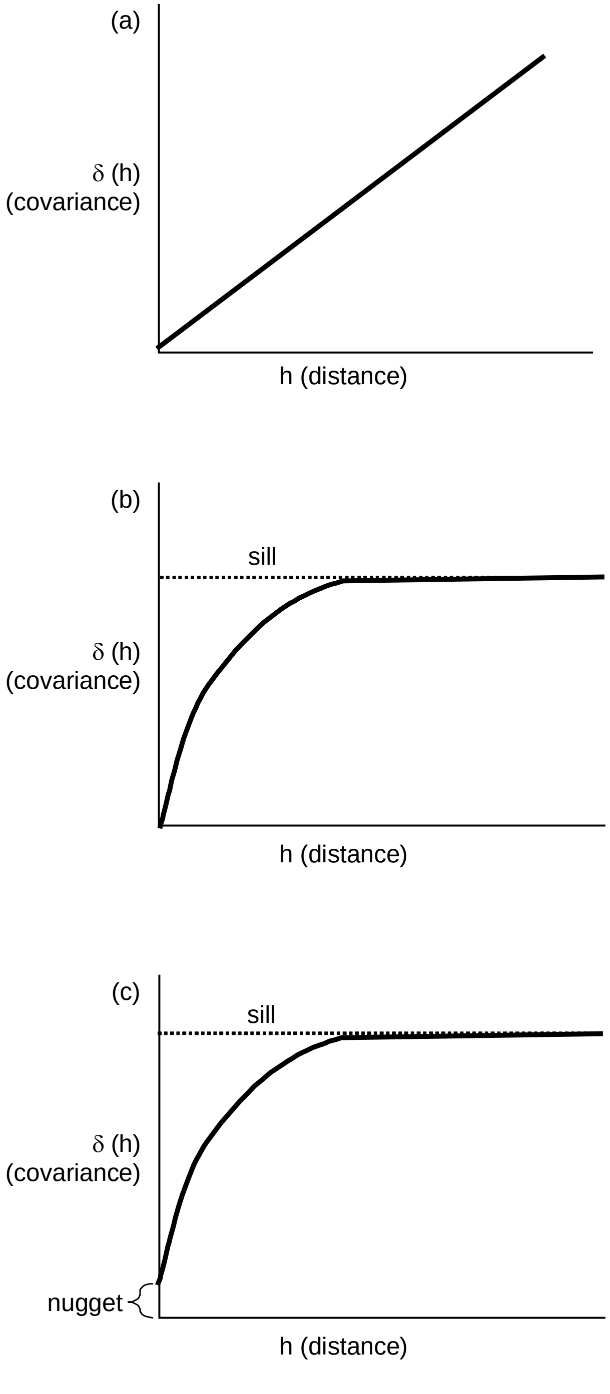


Figure R-1. Types of Semivariograms.

R-3.4. Any estimation procedure has an associated estimation variance. The special property of kriging is that it selects the set of weights that minimizes the estimation variance and produces the best linear unbiased estimator.

R-3.5. The assessment of uncertainty in geostatistics is highly quantitative; interpolated concentrations are estimated on the basis of an underlying model of correlation and variability. As such, the estimates themselves are directly linked to estimates of uncertainty. A predicted value may be expressed as a quantity plus or minus some quantity representing the uncertainty ($X \pm \varepsilon$), or the predicted value may be associated with a probability (X, $p = 0.9$). Specific methods of estimating the uncertainty are beyond the scope of this document; they are usually calculated using computer software.

R-3.6. Geostatistics can be used to evaluate and manage the uncertainty associated with remedial activities for a study area. Even with ample site characterization data (borings or wells), the boundaries of the treatment zone are imperfectly defined. Geostatistics allows us to evaluate the risk that the size, and, therefore, cost, of the remediation may be larger or smaller than expected. First, the site is characterized and adequate data are collected. Second, the data are transformed by assigning a value of 1 or 0 (indicator values), depending on whether the value is above or below, respectively, a given cleanup value or other criterion. Third, the transformed data are used to construct a variogram. Fourth, the variogram is modeled as previously described. This model is then used to perform kriging with the indicator values. The kriging estimates reflect a probability that the concentration at the points of estimation exceed the cleanup value or other standard. These kriging estimates can be contoured to define areas or volumes of material that have a certain likelihood of exceeding some cleanup value. The contour value is essentially the probability of exceedance. Last, the size of the area defined by different probabilities of exceedance can be determined and, using a unit cost or similar approach, a cost-versus-risk curve can be developed.

R-3.7. This can be used in programming money for the project, as a basis for negotiating cleanup levels with regulators, or to help determine if the cost and time of additional characterization work will be offset by less risk during construction. Alternatively, rather than transforming the data to ones and zeros, the actual values can be kriged, and the kriging variances can be used to determine prediction intervals for each estimated value. In the vicinity of the point estimate, these prediction intervals can be used to define the spread of potential values expected within a given probability. This assumes the data are normally distributed or have been transformed to be normally distributed.

R-4. Software for Geostatistics. There are a number of software applications to assist in geostatistical calculations. Two older applications developed by the U.S. Environmental Protection Agency (EPA) are GeoPack and Geo-EAS (EPA 600/4-88/033).

R-4.1. Repack conducts analysis of variability for one or more random functions. GeoPack includes basic statistics, such as mean, median, variance, standard deviation, skew, and kurtosis. The package also does regressions, distribution testing, and percentile calculations. Sample semivariograms, cross-semivariograms, or semivariograms for combined random functions for a two-dimensional, spatially dependent random function can also be determined. GeoPack includes ordinary kriging and co-kriging estimators in two dimensions, along with their associated estimation variance and the conditional probability that the value is greater than a user-specified cutoff level. Graphical tools include linear or logarithmic line plots, contour plots, and block (pixel) diagrams.

R-4.2. Geo-EAS was also developed by the EPA and is a collection of interactive software tools for doing two-dimensional geostatistical analyses of spatially distributed data. Programs are provided for data management, data transformations, univariate statistics, semivariogram analysis, cross-validation, kriging, contour mapping, post plots, and line-and-scatter graphs. The application is DOS-based.

R-4.3. A publicly available package of geostatistical software that is more comprehensive than these EPA packages is GSLIB, available at http://www.gslib.com. The DOS-executable freeware may be downloaded from this site. Alternatively, the software source code and a supporting textbook may also be purchased at the site for a nominal fee.

R-4.5. Commercial software for Windows, Sun, or Macintosh systems include WinGSLIB, Environmental Visualization System, and Groundwater Modeling System (GMS), which is currently available to all USACE, U.S. Department of Defense, EPA, and U.S. Department of Energy personnel.

R-5. Case Study: Geostatistical Analysis of Remediation by In Situ Ozonation.

R-5.1. Introduction. An application of geostatistics to environmental remediation will be explored in this case study. Three-dimensional kriging was used to support the Remedial Investigation/Feasibility Study, Remedial Action Plan, Confirmation Sampling, and Remedial Action Report (site closure) for a former manufactured gas plant (MGP) located in Long Beach, California. The former MGP operated from approximately 1901 to 1913 and produced gas from coal and crude oil feedstocks. The project was conducted pursuant to an agreement with the California Environmental Protection Agency Department of Toxic Substances Control under their Expedited Remedial Action Program. In-situ ozonation was used to lower levels of polycyclic aromatic hydrocarbons (PAHs) to meet the selected risk-based cleanup levels for this site. The kriging results played an important role in several estimation and decision processes, including:

R-5.1.1. Contouring the original distribution of PAH.

R-5.1.2. Defining the footprint and depths for the treatment zone.

R-5.1.3. Supporting decisions regarding placement for the ozone-injection well system.

R-5.1.4. Selecting quarterly monitoring locations for soil samples during the treatment process as well as for post-treatment confirmation samples.

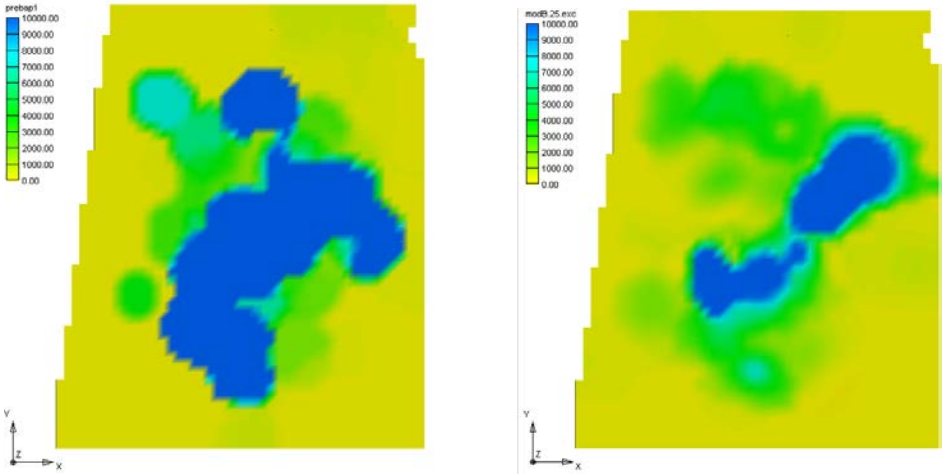
R-5.1.5. Contouring the final post-treatment distribution.

R-5.1.6. Estimating the site-wide exposure concentration used for risk assessment and site closure.

R-5.2. Post-Treatment Modeling. Quarterly monitoring results indicated substantial reductions in PAH levels early in the treatment, which began in 1998. However, within 2 years, monitoring indicated that the reductions had reached an asymptote, reflecting the diminishing return of continued ozonation and the recalcitrant nature of the residual PAHs. In 2000, confirmation samples were taken from random locations within the defined treatment zone. Kriging was then used to model the post-treatment spatial distribution of PAHs and compare it to the pre-treatment distribution (Figure R-2). Kriging uncertainty was estimated and used to determine whether cleanup goals had been met.

R-5.3. Reporting. The reporting of the kriging analysis was included in the Remedial Action Report as an appendix with an organization and level of detail consistent with guidelines given in “Standard Guide for the Contents of Geostatistical Site Investigation Report” (ASTM D5549-94e1). To enhance the practical value of this case study, the following parts of the ASTM outline are used below: software, data sources, exploratory analysis (and conceptualization), spatial continuity analysis, estimation, and uncertainty.

R-5.4. Software. The analysis was conducted using the three-dimensional kriging utilities of the GMS software mentioned in Paragraph R-4.



a. Pre-ozonation. b. Post-ozonation.

Figure R-2. Comparison of Krige-Interpolated Benzo(*a*)pyrene Concentrations before and after Treatment by In Situ Ozonation.

R-5.5. Data Sources. The variable of interest was benzo(a)pyrene equivalents (a weighted sum of carcinogenic PAHs in each sample). The first kriging analysis (pre-remediation) was conducted on sample data dispersed over approximately 75 soil borings primarily from the remedial investigation (RI) program that was completed in 1997. In contrast, the post-remediation kriging was conducted on a composite data set that consisted of 1997 RI samples that were outside the treatment area, together with the latest samples available for the treatment area taken in 2002. Thus, the post-remediation data set reflected the assumption that soil concentrations in the untreated areas were stable over time (reasonable for the PAHs involved) and, therefore, well represented by older data, while soil concentrations in the treatment zone were expected to change over time so that older samples were not included in the kriging analysis.

R-5.6. Exploratory Analysis and Site Conceptualization. The recommended Exploratory Analysis section in the ASTM guidelines is expanded here to be a conceptual discussion, deemed important for all sites, that considers all relevant qualitative and quantitative information about the site. The integration of these different types of information is crucial for explicitly identifying a conceptual model of the contamination distribution that will guide a number of assumptions and decisions throughout the analysis. Beyond the analytical sample results, such information includes topography, stratigraphy, observations made in boring logs, site history, and other qualitative and semi-quantitative information. For the MGP site, all examples from the above list were applied in some way during formulation of the geostatistical analysis. The following description of some of the qualitative information about the site is included before the transition into the exploratory data analysis.

R-5.6.1. Well-established site history provided engineering process information, as well as maps of potential source structures, that could be used to compare with the posted analytical results. An additional factor at the site is that its current condition includes an engineered soil levee along the Los Angeles River as well as soil fill set around large concrete supports for a bridge and on-ramp built across the site in 1953–1963 (subsequent to decommissioning of the MGP). Thus, the topography is quite varied and includes imported soil brought in to cover large parts of the site. Topography, native or fill, definitely influenced soil volumes and, therefore, had to be incorporated explicitly into the kriging estimation. Furthermore, the three-dimensional visualization of the topography and sample data (Figure R-3) indicated that spatial correlation occurred along a relatively level elevation rather than following the highs and lows of the present surface topography. (An approximate two-fold vertical exaggeration is used to aid the visualization of data points within a boring.) This is consistent with the expected pattern produced by an originally flat plant site. Because the subsequent mixing and earth movement are somewhat uncertain, the large volume of soil covering the former plant was sampled, along with the native soil, as part of the RI and was included in the site-wide model and calculations.

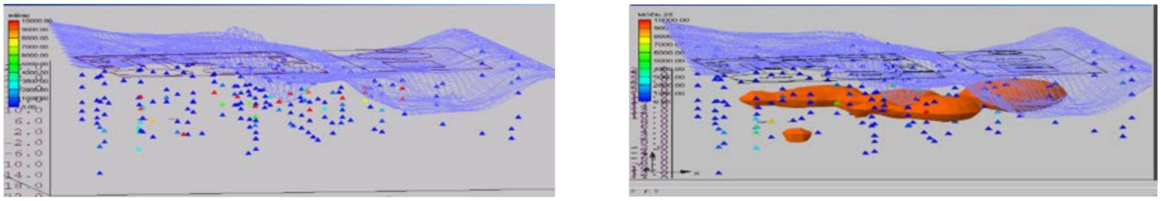


Figure R-3. Surface Topography with Benzo(*a*)pyrene Concentrations and a Kriged Isovolume.

R-5.6.2. The additional exploration of the analytical data, in the form of a histogram and descriptive statistics, indicated a high degree of skewness with suggestions of a composite of two different populations: one with low concentrations (i.e., “background”) that were found in the outlying areas, and the other with moderate-to-relatively high concentrations that still presumably reflected some varying amount of impact from the historical contamination (even after treatment). Samples in the outlying areas were sparser than in the central area, but still provided ample evidence to confirm the central positioning of the impacted soil in and around former MGP structures. Therefore, this potentially distinct population of low values was considered important to keep in the data set so that it would help define the outward extent of the residual contamination.

R-5.6.3. Including the low concentrations together with the more central data had the following implications.

R-5.6.3.1. Site Mean. The site-wide estimate of exposure concentration would reflect a site mean that included, in accordance with the defined site boundaries, both background and impacted volumes of soil. The site-wide mean to be calculated based on the kriging analysis would have contributions from both parts of the site in a manner that was “volume-weighted.” Given that any future redevelopment of the site would require the removal of the bridge support structures and intensive mixing of soil across the entire site, this site-wide mean was considered a realistic assumption for the conservative residential risk scenario.

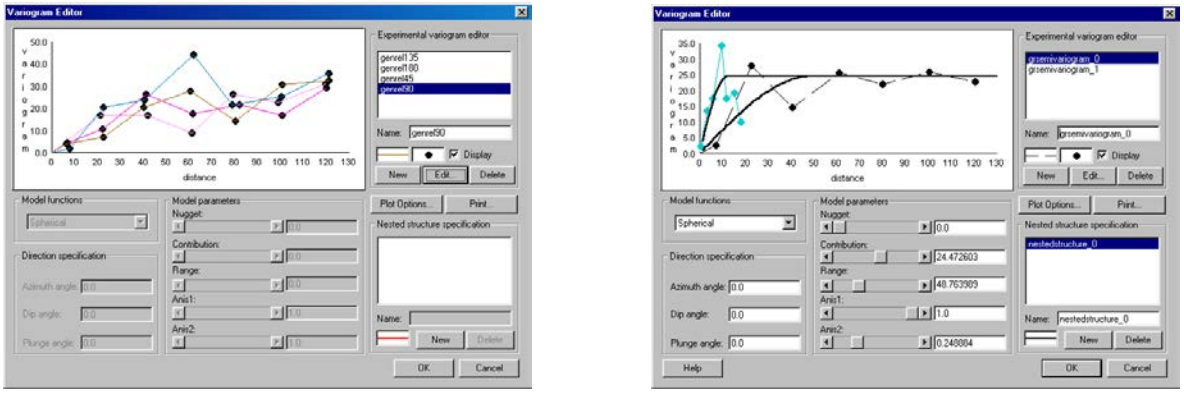
R-5.6.3.2. Spatial Pattern or Lateral Extent. The lower concentrations confirm site historical information regarding the “edges” of the impacted zone. Given this confirmation, the sparse outlying data can and should be supplemented with “soft data” to fill in areas of low data density and create a well-controlled boundary condition for the edges of the site. Such soft data, termed the “extended data set,” were added to the kriging for the estimation phase conducted after development of the variogram.

R-5.7. Spatial Continuity Analysis. It is reasonable to estimate soil concentrations across the site based on the underlying kriging assumption of spatial continuity. The fate and transport processes, involved in both the contamination and the ozone dispersion and effect, are presumably spatially continuous on some scale. Although soil structure and sample concentrations are notoriously variable or even discrete on a small scale, the resolution requirements implied by both risk assessment and remediation allow a broader focus. The view from risk assessment is one of exposure accumulated over time and space (a spatial average), and the view from remediation might be described as akin to the scoop size of a backhoe or some other scale useful for feasibility and cost estimation of the specific treatment. This larger scale of variability is more forgiving in the sense that an interpolation of the mean concentration in a cubic-yard block of soil has a lower uncertainty than an interpolation of any particular shovel-full of soil in the same block.

R-5.7.1. Therefore, the search for evidence and range of spatial continuity need not be a matter of finely tuned research for many sites although the level of rigor must be consistent with the site conceptualization. For example, one might argue that the outlying data, if they are truly background, may be a different population altogether than the central data, with different continuity ranges to be found by analyzing the two sets separately. On the other hand, there is no bright line around the site to delineate these two populations spatially (at least “a priori,” before the spatial analysis was done). More realistically, there is likely to be a gradient of soil impacted by some level of contamination and also some level of remediation, such that the net impact, or probability of impact, on the soil decreases with distance from the central area and individual injection wells. The spatial range of correlation that is defined for the kriging variogram should ideally be appropriate for this transition zone, as well as for the obvious central or outlying areas of the site. In other words, practicality points to the simplest assumptions that will “work.”

R-5.7.2. Spatial continuity was investigated on the entire post-remediation data set using a general relative variogram, which automatically adjusts for the proportional effect commonly found in contaminant concentration data and lognormal tending data in general. The variances calculated for a relative variogram were modified by dividing the group variances by the square of the local mean, which can be calculated in several ways. This improved the structure of the experimental variogram and, specifically for the case study data, allowed the modeler to observe lower relative variances (stronger correlations) at inter-point distances of about 5 to 10 feet (in the laterally direction), moderate variances at about 30 to 40 feet, and highest variances reaching a plateau at about 50 to 60 feet (see Figure R-4a).

R-5.7.3. Horizontal anisotropy was reviewed by limiting vertical and angular grouping parameters (depends on software package) to create different directional “horizontal variograms.” No horizontal anisotropy was present. However, the comparison between the horizontal variograms and vertical variograms, created by limiting horizontal grouping parameters, indicated that the vertical range of spatial correlation was approximately one-fourth that of the horizontal range. A spherical model variogram with vertical anisotropy was selected, with a horizontal range of 49 feet and a vertical range of 12.4 feet (Figure R-4b).



Variogram Editor

Experimental variogram editor: genrel135, genrel180, genrel45, genrel90

Name: genrel90

☑ Display

Model functions: Spherical

Model parameters: Nugget: 0.0; Contribution: 0.0; Range: 0.0; Anis1: 1.0; Anis2: 1.0

Direction specification: Azimuth angle: 0.0; Dip angle: 0.0; Plunge angle: 0.0

Nested structure specification

Name:

Variogram Editor

Experimental variogram editor: grsemivariogram_0, grsemivariogram_1

Name: grsemivariogram_0

☑ Display

Model functions: Spherical

Model parameters: Nugget: 0.0; Contribution: 24.472603; Range: 48.763989; Anis1: 1.0; Anis2: 0.240884

Direction specification: Azimuth angle: 0.0; Dip angle: 0.0; Plunge angle: 0.0

Nested structure specification: nestedstructure_0

Name: nestedstructure_0

a. Directional Horizontal Variograms.

b. Horizontal and Vertical Variograms.

Figure R-4. Two Sets of Experimental Variograms Used to Define the Krige Variogram Model.

R-5.8. Estimation. A concentration estimate was developed for every cell center of a three-dimensional grid with cell dimensions 6 by 6 by 3 feet. Cell dimensions were chosen to be consistent with earlier modeling work, but also were considered to provide an adequate balance between the resolution needs of risk and remediation resolution and the increased run-time and overall unwieldiness of denser grids. The three-dimensional contour map could be compared to the pre-remediation maps in plan view by layers, or by cross sections or rendered iso-volumes in GMS. The site-wide mean was then simply a matter of calculating the arithmetic average of all cell mid-points that were defined as “soil” (as opposed to “above ground”).

R-5.9. Uncertainty. Kriging standard error estimates are automatically produced for each cell at the time the concentration estimate is assigned. They reflect uncertainty in a particular cell estimate and cannot be used directly to estimate uncertainty for the site-wide mean, which is the standard error term required for a 95% upper confidence limit (95% UCL), i.e., the exposure concentration. The standard error for the site-wide mean was conservatively estimated by using the kriging error resulting when the variogram model was run on a new grid consisting of one large three-dimensional cell encompassing the entire site. The intuitive definition of this error term is that it represents the uncertainty implied by using the available 233 spatially correlated sample points to estimate the mean concentration of the entire block of soil containing the 233 correlated samples. As the site boundaries and especially the topography result in an irregularly shaped zone of soil within this large rectangular block, the “block type” of uncertainty results in an overestimate for the actual soil subzone within the block. This is because the block uncertainty reflects large regions of “air” that are not properly distinguished from soil, and these regions have no sample data and are relatively far from the nearest sample datum. This method was conservative but was considered reasonable for use in the risk assessment. Detailed discussions of the many uncertainty approaches for kriging can be found in Meyers (1997), which focuses on environmental contamination, and other geostatistical texts (e.g., Goovaerts, 1997). Thus, the 95% UCL on the mean was calculated as

$$95\%\ UCL = UE + 1.645\ KSE$$

where

UE = unbiased estimate of the mean (obtained from the high resolution model)

KSE = kriging standard error (conservative estimate)

1.645 = 95th percentile of the standard normal distribution

R-6. Conclusion.

R-6.1. Although several conservative analysis assumptions were built into the model and uncertainty formulation, the site-wide volume-weighted exposure concentration (95% UCL) was reduced by 37 to 58% compared to that calculated from the most commonly used non-spatial formulas identified in numerous risk assessment guidances (e.g., *t*-based, Land, bootstrap). The reduction in the exposure concentration came from the more rigorous use of spatial correlation and soil volume when kriging rather than the classical assumption that all sample points were identically distributed, i.e., without spatial correlation. Thus, the lower kriged exposure concentration was important in determining the attainment of risk-based cleanup goals.

R-6.2. The kriged model contours of the post-treatment spatial distribution allowed the visual comparison of the estimated pre- and post-remediation distributions, and were instrumental in concluding the effectiveness of in situ ozonation for this site.

APPENDIX S

Geochemical Trend Analysis

S-1. Introduction. An overview of the "geochemical" approach is presented from a statistical perspective via illustrations, and existing geochemical guidance (primarily from the Navy) is supplemented. The geochemical approach is an effective strategy for distinguishing anthropogenic from naturally occurring metal concentrations, particularly when it is used with traditional quantitative statistical evaluations. The approach often identifies naturally occurring metal concentrations that are erroneously identified as site- related by traditional evaluations (i.e., comparisons of study area metal concentrations to background 95% UTLs). The geochemical approach can not only be used to determine whether a study area has been impacted by anthropogenic metal contamination but can also identify the individual sampling locations that are suspected to possess the elevated metal concentrations.

S-1.1. Although the geochemical approach is typically extremely useful, the limitations of the approach should be noted. Its primary disadvantage is that it is subjective because it is predominately qualitative. In particular, decision errors are not quantified and well-defined criteria for distinguishing native from anthropogenic metal concentrations are not specified. In addition, although the approach distinguishes anthropogenic metal contamination from naturally occurring concentrations, it does not distinguish site-related contamination from non-site-related anthropogenic metal contamination. In other words, elevated contamination relative to background identified by the geochemical approach may be consistent with anthropogenic background. Statistical comparisons using a background study area would typically be needed to distinguish site-related contamination from total background metal concentrations (from anthropogenic and non-anthropogenic sources). Lastly, an additional limitation of the approach is that it implicitly assumes that, at most, only a portion of the site has been impacted by anthropogenic metal releases. This assumption is typically reasonable but can be violated if the study area is too small (i.e., is predominately limited to a "hot spot").

S-1.2. Geochemical evaluations may be categorized as "association" and "enrichment" analyses. Both are qualitative strategies used to distinguish anthropogenic from naturally occurring metal concentrations and rely upon the assumption that metal releases from waste handling activities impact only a portion of the study area. Geo-chemical "association" analysis primarily uses scatter plots to distinguish anthropogenic from naturally occurring metal concentrations. The approach exploits and relies upon the ability to observe correlations between different naturally occurring metals, while geochemical "enrichment" analysis primarily uses probability plots to accomplish this objective. Typically (for both geochemical approaches), at least 20 samples are collected for some environmental medium of interest at the study area (i.e., surface soils or groundwater that has been potentially impacted by metal contamination) and the samples are analyzed for TAL (target analyte list) metals (i.e., the set of 23 metals listed in the Contract Laboratory Program Statement of Work). Because metals such as Al, Mg, Ca, and Fe are major components of naturally occurring minerals in rocks and soils in the earth's crust, these metals are typically considered to be non-site related.

S-1.3. When geochemical association analyses are done, correlations between suspected site-related metals (e.g., Cd, Pb, and Cu) and non-site related metals (e.g., Al, Fe, or Ca) are investigated by generating scatter plots. Typically, the concentrations of some potential site-related metal are plotted on the y-axis and the corresponding concentrations of some non-site-related metal are plotted on the x-axis. A strong correlation suggests that detected metal concentrations are native rather than a result of site-related waste handling activities. Metal concentrations that are not consistent with the correlations in the scatter plots appear as “anomalies” or “outliers” that are attributed to anthropogenic contamination. When geochemical enrichment analysis is performed, probability plots are generated. Native metal concentrations give rise to continuous monotonic curves (i.e., straight lines). An abrupt increase in the slope of a curve, appearing as an inflection point in the upper portion of the curve, indicates anthropogenic contamination.

S-1.4. The strategies used to select the particular native metals of interest are beyond the scope of this document, which focuses upon only the statistical evaluation of the data once the metals of interest have been selected. The metals and the correlations of interest will depend on the nature of the environmental population being sampled. Native metals concentrations in soils and sediment depend on factors such as the nature of the parent rocks and component minerals, and organic material content. Metal concentrations tend to be directly proportional to total organic carbon and inversely proportional to particle size. Dissolved metal concentrations in groundwater tend to be greater at low pH and reducing conditions. It should be noted that metals usually exist as anions (negatively charged species) and cations (positively charged species) in environmental media such as groundwater, soil, and sediments. For example, metals such as As, Sb, Se, V, and Mo tend to form anionic species (i.e., containing oxygen atoms); metals such as Ba, Cu, Pb, Ni, and Zn tend to form cations, while certain metals such Cr form either as cationic or anionic species. At neutral pH, clays, which typically contain Al, possess strong negative surface charges that attract cationic metals such as Cu, Zn, and Pb. Therefore, for soils rich in clay or groundwater containing suspended clay particles, Al will often be strongly correlated with cationic metals. Similarly, at neutral pH, environmental matrices containing iron oxides and iron oxyhydroxides possess positive surface charges that attract anionic metal species.

S-2. Geochemical Association Approach. To illustrate the geochemical association approach, assume that soils at some study area contain significant concentrations of native Fe and the area is suspected to have been impacted by site-related Pb contamination. The concentration of Pb in each sample is plotted against the corresponding concentration of Fe to generate a “Pb-Fe” scatter plot for the study area (i.e., as discussed in Paragraph J-9). When a scatter plot is generated for a geochemical evaluation, the x-axis is usually the concentration of the non-site-related metal (Fe), but this is merely a convention (e.g., a comparable scatter plot may be generated if the y-axis were the concentration of the non-site related metal). Also note that when the scatter plot is produced, the values for the X variable and those for the Y variables are not ordered prior to plotting the data, rather a set of paired measurements (x_i, y_i), where i = 1, 2, …, n (n denotes the number of environmental samples) is plotted. A strong positive correlation between naturally occurring concentrations of Fe and Pb (i.e., where the concentration of Pb tends to increase as the concentration of Fe increases) would suggest that Pb is not an anthropogenic contaminant. Figure S-1 is an example of a Fe-Pb scatter plot.

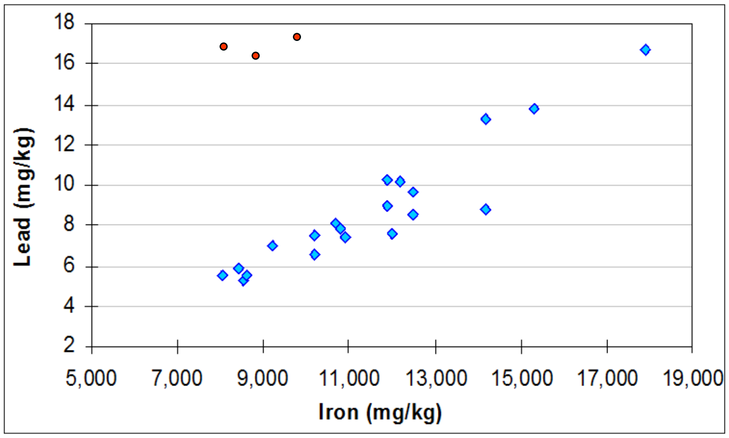


Figure S-1. Scatter Plot of Pb and Fe. Copyright 2004 From "Identifying Metals Contamination In Soil: A Geochemical Approach," Soil & Sediment Contamination, Vol. 13, No. 1, pp. 1–16, by Myers, J. and K. Thorbjornsen. Reproduced by permission of Talyor & Francis Group, LLC.

S-2.1. The relatively strong linear relationship between Pb and Fe for the points that appear as blue diamonds suggests that these samples contain only native concentrations of Pb and Fe. Samples containing Pb in excess of naturally occurring concentrations appear as "outlying" points (e.g., the three red circles) above the linear trend (the blue diamonds), suggesting that these samples contain anthropogenic Pb contamination.

S-2.2. Two major advantages of the geochemical approach relative to classic statistical approaches are immediately apparent. A background study area (and the expense associated with doing a separate background study) is not required to identify study area concentrations that are elevated relative to native metal concentrations. Furthermore, the approach readily identifies the samples (locations) suspected to contain the elevated metal concentrations. Classic statistical evaluations do not readily provide this information. (Because classic statistical evaluations rely upon the assumption that samples are independent of one another, the presence of a correlation or contamination "pattern" would violate this assumption and compromise the validity of the evaluation.) For example, a typical statistical approach would entail comparing the mean concentration of Pb at the site study area to the mean concentration of Pb at a background study area. Although the evaluation may indicate that the mean site Pb concentration is statistically greater than the mean background Pb concentration, the evaluation itself would not (at least directly) identify the sampling locations associated the elevated lead concentrations (though a geostatistical approach could potentially evaluate contamination that is spatially correlated).

S-2.3. It should also be noted that, although background data are not required to perform geochemical evaluations, background data can be plotted with site data to determine if site metals are elevated relative to native concentrations. This is illustrated in Figures S-2 and S-3.

S-2.4. In Figure S-2, the Cu surface soil samples (blue non-shaded triangles) generally plot above the background samples (green circles). Similarly, in Figure S-3, Pb surface soil samples (blue non-shaded triangles) plot above the background samples (green circles). This suggests that the site has been contaminated by both Pb and Cu. These plots were generated from soil samples collected from an artillery firing range, where Cu and Pb are frequently potential contaminants of concern. The scatter plots also indicate that Pb and Cu in the site surface soils are elevated relative to the subsurface soils, which, given the nature of the site, is consistent with the manner in which one would expect site-related contamination to be spatially distributed.

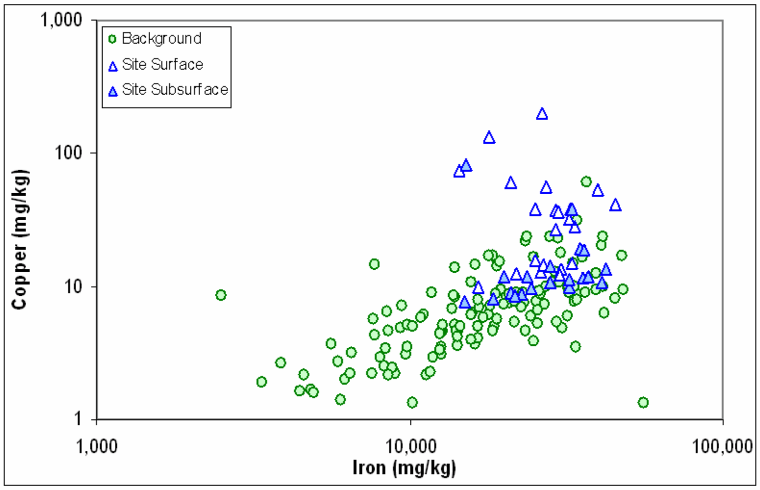


Figure S-2. Log Scale Cu-Fe Scatter Plots of Site and Background Soil Samples. Figure provided by J. Myers of Shaw Environmental, Inc., Knoxville, TN.

S-2.5. An additional advantage of the geochemical approach is that multiple scatter plots between different metals (i.e., using site or a combination of site and background data) can potentially be used to determine whether or not a site has been contaminated by metals. In this example, the anthropogenic Cu and Pb contamination identified in the Cu-Fe and Pb-Mn scatter plots, respectively, can be further evaluated by generating a scatter plot for Pb and Cu, as shown in Figure S-4. The moderate to strong correlation between Cu and Pb for the site surface soil samples but the poor correlation between Pb and Cu for the background samples suggests that the Cu and Pb are site-related contaminants from a common anthropogenic source.

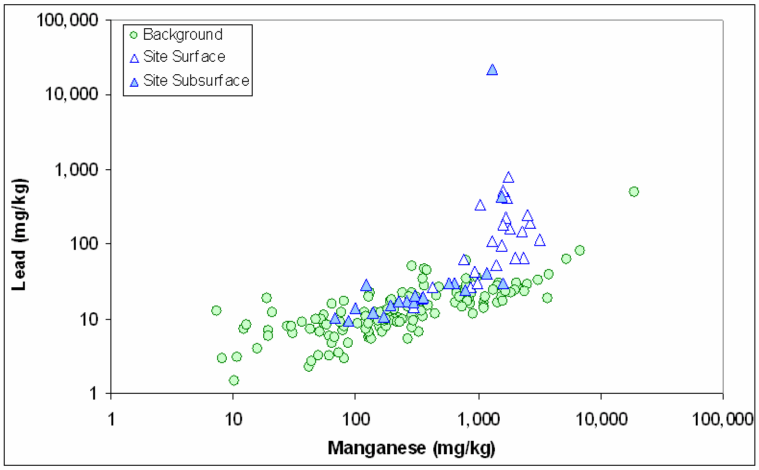


Figure S-3. Log Scale Pb-Mn Scatter Plot for Site and Background Soil Samples. Figure provided by J. Myers of Shaw Environmental, Inc., Knoxville, TN.

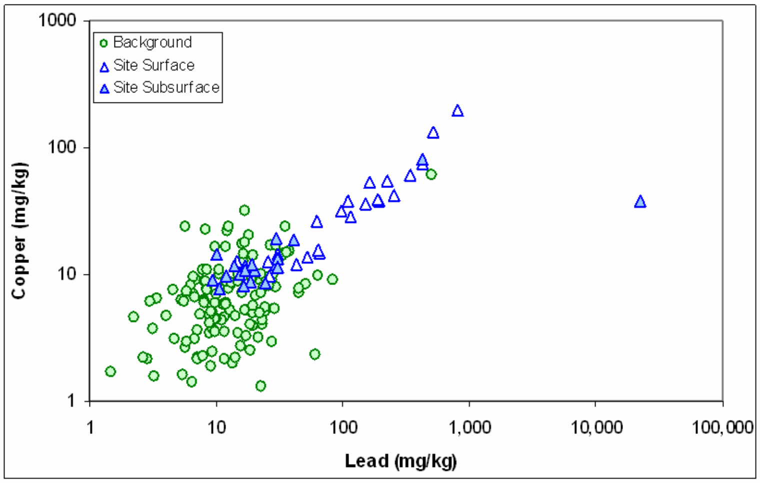


Figure S-4. Log Scale Cu-Pb Scatter Plot of Background and Site soils. Figure provided by J. Myers of Shaw Environmental, Inc., Knoxville, TN.

S-2.6. As stated previously, the primary disadvantage of the geochemical approach is that it is predominately qualitative and, therefore, subjective. The degree of correlation that is required to conclude the study area has not been affected by anthropogenic contamination and what constitutes an “outlier” when a correlation is observed is typically is not well defined (i.e., quantitatively criteria are not specified). To illustrate, consider the As-Fe scatter plot presented below in Figure S-5.

S-2.7. There appears to be a large of amount of dispersion in the scatter plot shown in Figure S-5. A qualitative visual evaluation of this plot does not clearly indicate whether or not As and Fe are strongly correlated with one another. However, as illustrated in Figure S-6, the same scatter plot could potentially be interpreted in a different way: Arsenic concentrations less than about 4 mg/kg could be viewed as strongly correlated with Fe (as shown by the red line in Figure S-6), and the As concentrations larger than 4 mg/kg (i.e., the set of circled points) could be interpreted as anthropogenic contamination. Unlike classical statistical strategies that are used to distinguish anthropogenic contamination from background values, decision errors for geochemical evaluations are not quantifiable. As geochemical evaluations are subjective, they can produce erroneous conclusions and are more vulnerable to challenge (e.g., by regulators) than quantitative statistical approaches.

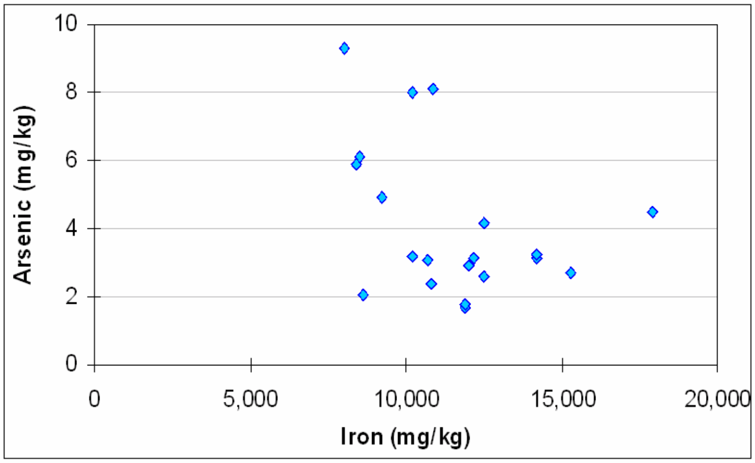


Figure S-5. As-Fe scatter plot with a large amount of scatter. Copyright 2004 From "Identifying Metals Contamination In Soil: A Geochemical Approach," Soil & Sediment Contamination, Vol. 13, No. 1, pp. 1–16, by Myers, J. and K. Thorbjornsen. Reproduced by permission of Talyor & Francis Group, LLC.

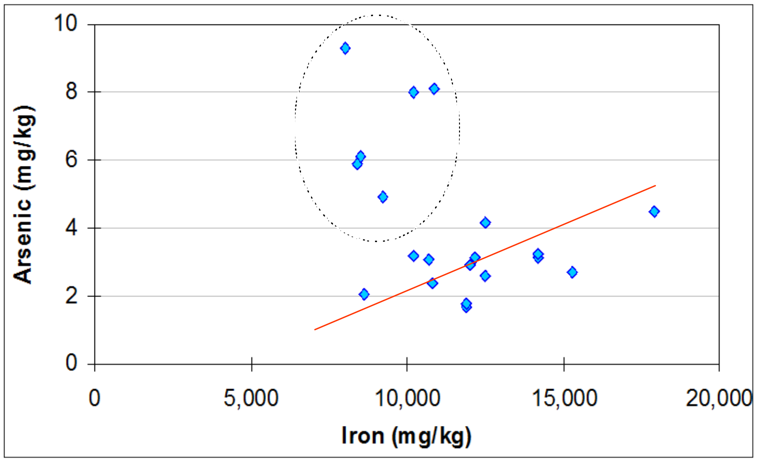


Figure S-6. Misidentified trends for the scatter plot in Figure S-5. Copyright 2004 From "Identifying Metals Contamination In Soil: A Geochemical Approach," Soil & Sediment Contamination, Vol. 13, No. 1, pp. 1–16, by Myers, J. and K. Thorbjornsen. Reproduced by permission of Talyor & Francis Group, LLC.

S-2.8. However, the As results in Figure S-5 are probably naturally occurring. As shown in Figure S-7, a scatter plot of As versus the ratio Ln(As/Fe) exhibits a fairly strong linear relationship, suggesting that the As is natural.

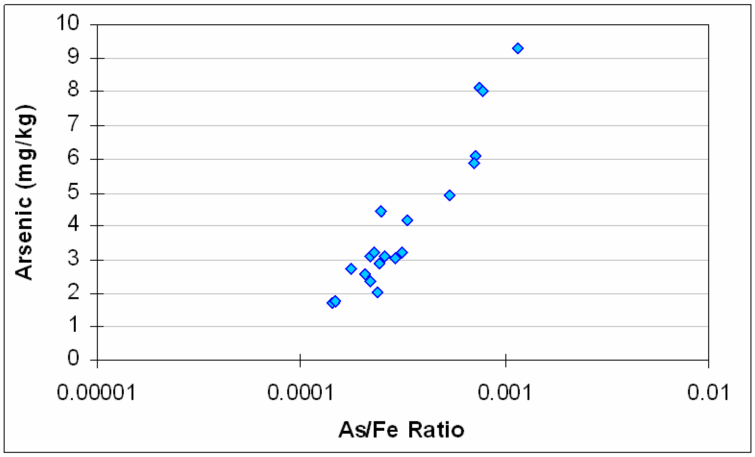


Figure S-7. Scatter Plot of As and Logarithm of As/Fe Using the Data set Plotted in Figure S-5. Copyright 2004 From "Identifying Metals Contamination In Soil: A Geochemical Approach," Soil & Sediment Contamination, Vol. 13, No. 1, pp. 1–16, by Myers, J. and K. Thorbjornsen. Reproduced by permission of Talyor & Francis Group, LLC.

S-2.9. The scatter plots presented above were generated using soils data, but similar geochemical association analyses may also be conducted for groundwater. Some scatter plots using log rather than linear scales for the x- and y-axes are presented below for groundwater data.

S-2.10. There is a relative good correlation between Al and Fe in Figure S-8, which suggests that both metals are non-site-related. The correlation between As and Fe in Figure S-9 suggests that As is not a site-related contaminant.

S-2.11. The scatter plots may also be used to examine the relationship between filtered and unfiltered samples, as well as between metal concentrations and parameters such as turbidity and oxidation-reduction potential (e.g., in single monitoring well over time or for a set of monitoring wells). Figure S-10 illustrates the relationship between filtered and unfiltered samples analyzed for Cr. There is an apparent linear relationship between the concentration of Cr in unfiltered groundwater and the ratio of filtered to unfiltered Cr, which could indicate naturally occurring Cr in suspended particles from the surrounding soils.

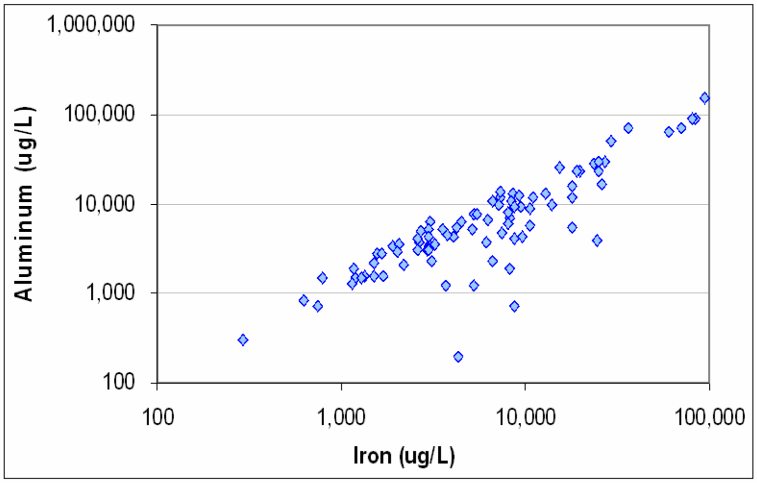


Figure S-8. Al-Fe log-scale Scatter Plot for a Set of Groundwater Monitoring Wells. Figure provided by J. Myers of Shaw Environmental, Inc., Knoxville, TN.

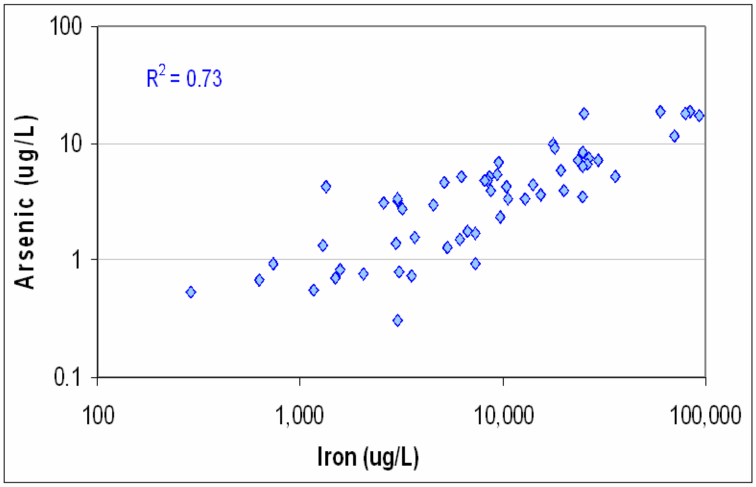


Figure S-9. Log-scale As-Fe Scatter Plot Using Fe Groundwater Data for Figure S-8. Figure provided by J. Myers of Shaw Environmental, Inc., Knoxville, TN.

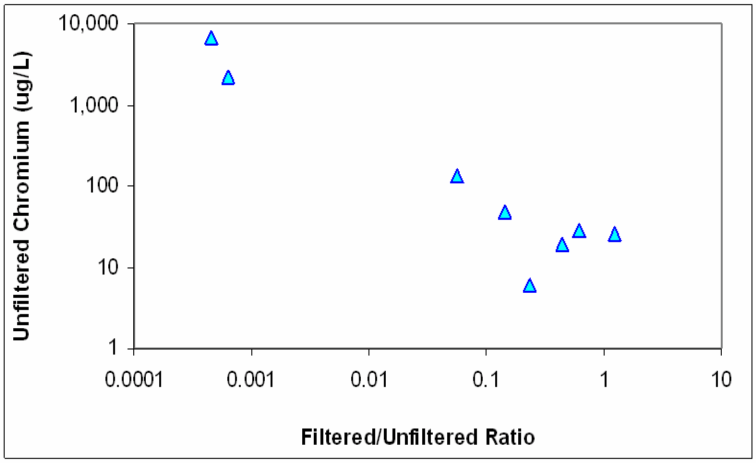


Figure S-10. Log-scale Scatter Plot of Filtered and Unfiltered Groundwater Analyzed for Cr. Figure provided by J. Myers of Shaw Environmental, Inc., Knoxville, TN.

S-3. <u>Geochemical Enrichment Analysis</u>**.** Geochemical enrichment analysis entails constructing quantile plots or normal probably plots (e.g., as discussed in Appendix J). To construct a quantile plot, the values of some variable are ordered from smallest to largest and the percentage or faction of the values less than or equal to each data point is then calculated. The measured values are then plotted on one axis (y-axis) and the corresponding percentages or proportions are plotted on the remaining axis x-axis). The approach is so named because the measured variable being plotted is called an "enrichment factor." An enrichment factor is calculated from an equation of the form:

$$Y' = (C_M / C_X)_{Site} / \mu_{Parent\ Rock} \quad .$$

S-3.1. The quantity $(C_M / C_X)_{Site}$ is the concentration of some site related metal (e.g., Cr) C_M divided or "normalized" by the corresponding concentration of some non-site- related metal (e.g., Al) C_X. The term $\mu_{Parent\ Rock}$ is the true mean concentration of (C_M / C_X) concentration in the "parent rock" (i.e., the rock from which the site soil was geologically derived) and is typically obtained from the literature. However, as this term is simply a constant, it does not alter the shape of the quantile plots and is unnecessary for their evaluation. Quantile plots may be generated using the ratios

$$Y = (C_M / C_X)_{Site}$$

or the logarithms of these ratios

$$Ln(Y) = Ln\{(C_M / C_X)_{Site}\} \quad .$$

S-3.2. The quantile plot is evaluated for trends indicative of naturally occurring metal concentrations and "deviations" that indicate anthropogenic contamination. Because environmental data are frequently normal or lognormal, it is usually convenient to construct normal probably plots for Y or Ln(Y) (i.e., the values of $(C_m / C_x)_{Site}$ are plotted against the corresponding quantiles of a standard normal distribution or their associated probabilities). For normally distributed data, "deviations" appear as "breaks" in a straight line. This is illustrated in Figure S-11.

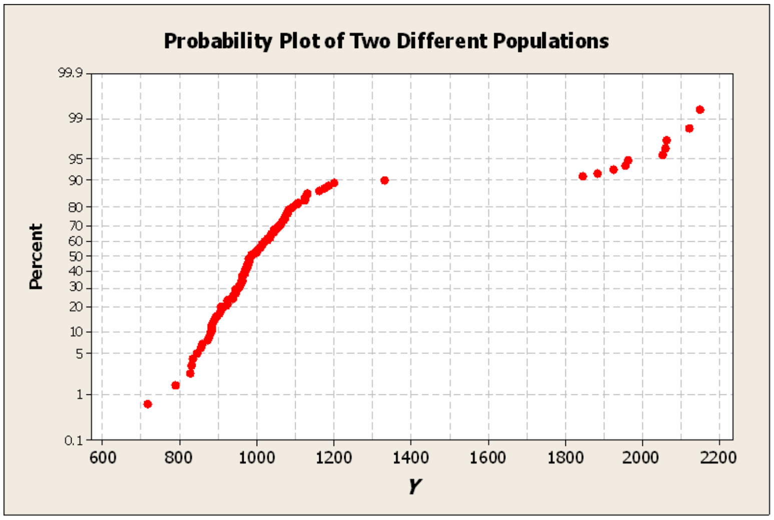


Figure S-11. Probability Plot of Y = $(C_M/C_X)_{Site}$ When a Portion of the Study Area has been Heavily Impacted by Anthropogenic Contamination.

S-3.3. The plot is predominately linear from about 700 to 1300, where there appears to be either a "break" or inflection point in the graph. After this region, the graph is essentially linear from about 1800 to 2200. The linear portion of the plot from 700 to 1300 would be attributed to native background concentrations and the values greater than about 1300 would be attributed to anthropogenic contamination. It should also be noted that the probably plots may contain more than one inflection point. Multiple populations (i.e., differences in concentration between background soils, surface soils, and subsurface soils) will potentially give rise to multiple inflection points. Ideally, the total number of inflection points plus one will be equal the number of different populations.

S-3.4. There are two apparent inflection points for the probability plot in Figure S-12 (one near 120 and one near 180), which suggests that there are three distinct populations. For example, there may be a background data set and two different concentration regions for site-related waste handling activities, or there may be two distinct background data sets and one data set for sampling locations impacted by anthropogenic contamination. However, the identification of the background "trend" and the "deviations" are subjective components of the evaluation. The value at which the "break" or inflection point occurs cannot be precisely determined, and accuracy decreases as the variability increases and the average native concentrations approaches the average concentrations of anthropogenic contamination.

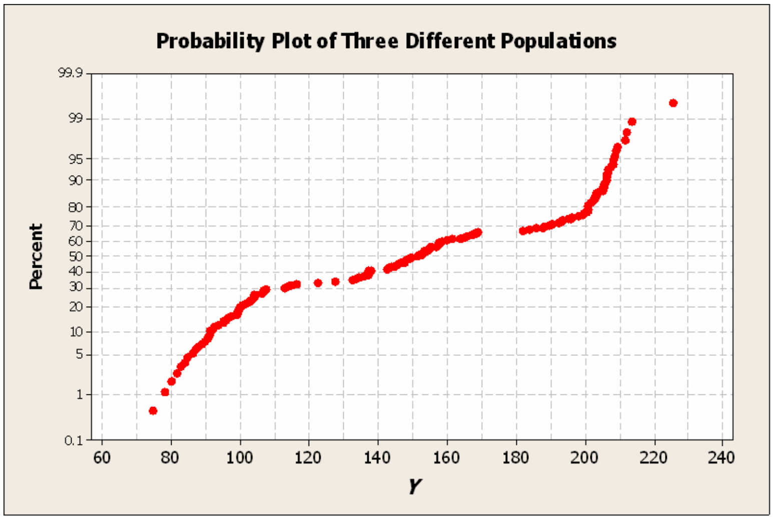


Figure S-12. Probability Plot of Y = (C_M/C_X) Site for Three Different Populations.

S-3.5. Two different known data sets were actually combined to produce the plot in Figure S-13. The "background data set" consisted of 100 points from a normally distributed population with a mean of 1000 and standard deviation of 100. The second set, which represents the anthropogenic contamination, consisted of 10 points from a normally distributed population with a mean of 2000 and a standard deviation of 100. As the difference between the means is large, an inflection point can be easily obtained from the probability plot in Figure S-13. However, a very different probability plot would result if the means of the two data sets were more similar. Consider the probability plot that would have been produced by combining the following data sets: i) a "background" data set, consisting of 100 points from a normally distributed "background" population with a mean of 1000 and standard deviation of 200, and ii) a "site" data set, consisting of 10 points (representing the anthropogenic contamination) from a normally distributed population with mean of 1300 and standard deviation of 200.

S-3.5. An inflection point is not apparent in the probably plot though the plot contains 10 data points from a population with a mean that is significantly greater than the background mean. Descriptive statistics for the two data sets are presented below:

Variable	**Mean**	**Std. Dev.**	**Minimum**	**Maximum**
$Y_{Background}$	1004.3	212.2	548.9	1592.3
Y_{Site}	1306.6	211.6	837.9	1512.9

S-3.6. Assuming that the background areas are known, a two-sample Student's t- test could show that there is a significant difference between the means for the "background" and "site" data sets at well over the 95% level of confidence. Unlike the geochemical approach, this test would conclude that the "site" is elevated relative to "background." As in the geochemical association approach, the qualitative nature of enrichment factor approach can produce decision errors. Geochemical evaluations should typically be done with quantitative statistical evaluations to determine whether or not a study area has been impacted by metal contamination.

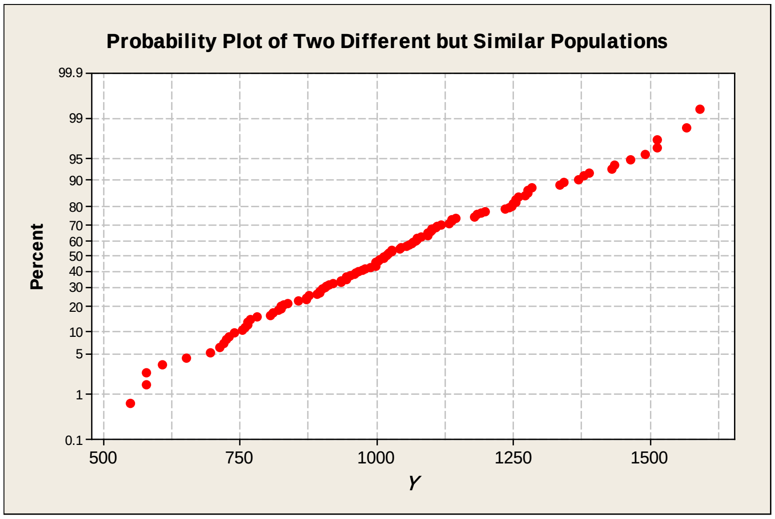


Figure S-13. Probability Plot of Y = (C_M/C_X) Site When a Portion of the Study Area has been Slightly Impacted by Anthropogenic Contamination.

S-4. Recommendation for Performing Geochemical Evaluations. Relatively detailed guidance for evaluating background concentrations using classic statistical as well as geochemical evaluations is available from the Navy for soil, sediment, and groundwater at the following web link: http://web.ead.anl.gov/ecorisk/related/. However, some modifications to the Navy's approach are recommended as listed below.

S-4.1. In the Navy guidance, Ordinary Least Squares (OLS) (linear regression) is used to evaluate geochemical relationships (e.g., correlation), outliers that present contamination, and is used to estimate background concentrations. It is recommended that OLS calculations not be performed. The underlying assumptions required to perform linear regression of typically violated (as discussed in Paragraph P-4).

S-4.1.1. As discussed in Appendix P, when a regression line of the form $Y = b_1 X + b_0$ is calculated, it is being assumed that X is an "independent" variable that possesses negligible uncertainty relative to the "dependent" variable Y. A change in X produces "explained" variation in Y; the "unexplained" variation is attributable to random error associated with the measurement of Y alone. However, this assumption is routinely violated for geochemical evaluations. In the Navy's guidance, non-site-related metals such as Al and Fe are plotted on the x-axis and potential site-related metals are plotted on the y-axis, but this is merely a convention. The variables X and Y are both measured quantities possessing comparable levels of uncertainty. In this context, there is no a prior justification for treating the two variables differently. Furthermore, other underlying assumptions required for regression fits are often (but not necessarily) violated (e.g., the residuals must be normally distributed and the variance cannot be a function of X or Y).

S-4.1.2. The violation of the underlying assumptions required to calculate the regression lines can produce erroneous conclusions. For example, when regression lines are calculated, the Navy guidance quantifies their certainty to calculate predication intervals, which are used to identify outliers indicative of anthropogenic contamination. (Points that lie outside the prediction intervals are suspected to be elevated relative to native concentrations.) However, when the assumptions required for the regression lines are violated, the prediction intervals will not necessary be valid, which may result in incorrect decisions.

S-4.2. Geochemical evaluations should focus (at least initially) on correlation rather than OLS regression. A correlation coefficient is a measure of the degree of association between two variables. Unlike regression, it does not require a “dependent” and “independent” variable. Three common measures of correlation are Pearson’s r, Kendal’s tau, and Spearman’s rho (refer to Appendix O). However, Pearson’s r is recommended only to screen the results for correlations (e.g., to generate the correlation matrix in Table 3-1 of the Navy’s soil guidance).

S-4.2.1. Pearson’s r measures only linear associations; is not appropriate when the data are not normal (a bivariate normal distribution is actually required), and is not invariant under logarithm transformations (e.g., Pearson’s r calculated for an X-Y scatter plot will differ from that calculated for a Ln(X)-Ln(Y) scatter plot). Furthermore, it is not appropriate when a significant number of non-detects are reported (i.e., not robust to data censoring). In contrast, Kendal’s tau and Spearman’s rho are non-parametric correlation coefficients (i.e., normality is not required) that measure the degree of association for monotonic (linear and non-linear) relationships. They are invariant with respect to monotonic transformation, such as logarithm transformation, and are relatively robust to data censoring.

S-4.2.2. A statistical hypothesis test should be performed for a correlation coefficient calculated for two sets of measured variables (metals), X and Y, to determine if it is statistically different from zero at the 95 or 99% level of confidence. If the correlation coefficient is not statistically different from zero, there is insufficient evidence to conclude that two variables (metals) are correlated with one another. If the coefficient is statistically different from zero, then we may conclude that some degree of associate exists. Unfortunately, there is no quantitative criterion for the degree of association. Two metals may exhibit a statistically significant correlation, but the degree of correlation may be so weak that it is not of practical importance. However, some “rule-of-thumb” guidance for the degree of correlation is presented in Paragraph O-2. It is recommended that at least a weak to moderate relationship be required for geochemical associations.

S-4.2.3. When non-detects are reported (especially when the non-detects are reported at different detection limits), it is recommended that correlation be evaluated using Kendal’s τ-$_b$:Kendal’s τ-$_b$ would typically be calculated using statistical software and is essentiallyKendal’s tau adjusted for tied values (see Appendix O).

S-4.3. A Kendal-Theil or “line of organic correlation” (LOC) should be plotted with scatter plots to help identify linear relationships (refer to Appendix P). A Kendal-Theil line passes through the medians of both variables X and Y that are linearly related. The slope of the Kendal-Theil line is not significantly different from zero if Kendal’s tau is not significantly different from zero. Unlike the least-squares regression line, the Kendal- Theil line is non-parametric and is relatively robust to outliers and censored data. The calculation of a LOC constitutes an alternative parametric approach to examine a linear relationship that would be more appropriate than OLS. A LOC is appropriate to evaluate linear relationships for the geochemical approach because the uncertainty associated with both sets of metal measurements is taken into account. The LOC is calculated in a similar manner as OLS lines, but the X and Y variables are treated in the same manner (i.e., the approach does not require “dependent” and “independent” variables). The same LOC will be obtained whether Y is plotted against X or X is plotted against Y.

S-4.4. The Navy guidance recommends that only Ln(X)-Ln(Y) scatter plots be generated. However, X-Y (or Ln(X)-Y, and X-Ln(Y)) scatter plots can also be generated and may be useful for identifying associations between variables, as shown by the X-Y scatter plot in Figure S-1. Associations can also be identified by generating scatter plots of the form: “X versus X/Y” (e.g., where X denotes the concentration of a potential site- related metal and X/Y is the ratio of the metal to a non-site-related metal concentrations). A linear relationship between X and X/Y implies that a linear relationship will be obtained when Y is plotted against Ln(X). (If Y is proportional to Ln(X), then the first derivative dY/dX is proportional to 1/X and dX/dY is proportional to X.)

S-4.5. The Navy’s groundwater guidance document does not promote the geochemical evaluations presented for soils and sediments in the Navy’s soil and sediment background guidance documents. The geochemical evaluations for soils and sediments can substantively be applied to groundwater, as shown by groundwater scatter plots presented above.

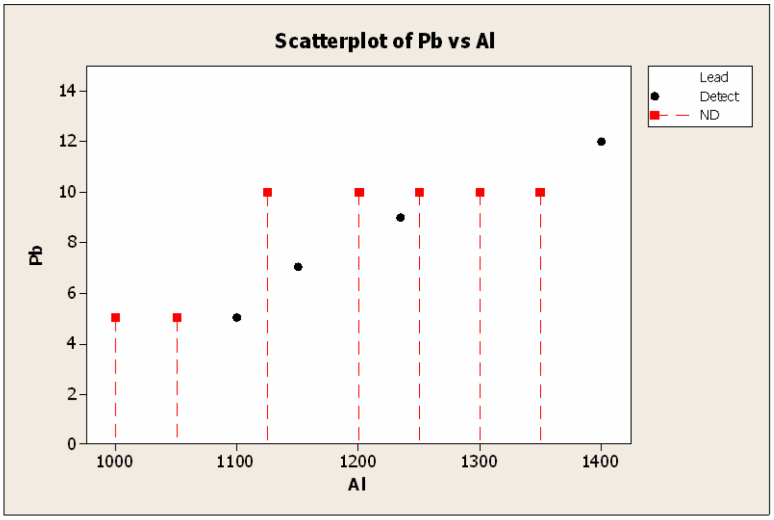


Figure S-14. Scatter plot for censored data.

S-4.6. For the geochemical enrichment approach, it is recommended that both the ratios $(C_m/C_x)_{Site}$ and the logarithms of the ratios be plotted to identify trends characteristic of anthropogenic contamination. The normalization factor $(C_m/C_x)_{Parent\ Rock}$ is not required and may be omitted if convenient to do so.

S-4.7. Censored data (non-detects) should be included in scatter plots for the geochemical association analyses when only one of the variables is censored. The uncensored variable (which would typically be a non-site-related metal such as Fe or Al) should be plotted along the x-axis and the censored variable (the suspected site-related metal) should be plotted on the y-axis. To illustrate, a Pb and Al scatter plot is presented in Figure S-14 for a small data set. The black circles represent detected results and the red squares are the reporting limits for non-detects. The dashed lines indicate that the actual Pb concentration lies somewhere between the reporting limit and the zero. One the basis of the detected results alone, there appears to be a strong correlation between Pb and Al. However, the correlations appears to be rather weak when the non-detects are also plotted.

GLOSSARY

Acronym	Definition
%R	Percent recovery
2-D	Two-dimensional
3-D	Three-dimensional
ACLs	Alternate concentration limits
ANOVA	Analysis of variance
ARARs	Applicable or Relevant and Appropriate Requirements
ASAP	Adaptive Sampling and Analysis Program
ASTM	American Society for Testing and Materials
CERCLA	Comprehensive Emergency Response, Compensation, and Liability Act
CFR	Code of Federal Regulations
CI	Confidence interval
COPCs	Contaminants of potential concern
CTE	Central tendency exposure
CUSUM	Cumulative summation
CV	Coefficient of variation
DCA	Dichloroethane
DDT	Dichloro-diphenyl-trichloroethene
DEFT	Decision error feasibility trial
df	Degrees of freedom
DLs	Detection limits
DO	Dissolved oxygen
DQI	Data quality indicator
DQO	Data quality objectives
EPA	U.S. Environmental Protection Agency
EPCs	Exposure point concentrations
EQL	Estimated quantitation limit
FSP	Field sampling plan
Geo-EAS	Geostatistical Environmental Assessment Software
GIS	Geographic Information System
GPS	Global Positioning System
HRS	Hazard ranking system
HTRW	Hazardous, toxic, and radioactive waste
IAA	Immunoassay analysis
ICV	Initial calibration verification
IDL	Instrument detection limit

IDW	Inverse distance weighted
IQR	Interquartile range
K-S	Kolmogorov-Smirnov
Lc	Critical level
LCL	Lower confidence limit
LD	Limit of detection
LS	Least squares
LSD	Least significant difference
MCLs	Maximum contaminant levels
MDL	Method detection limit
MQL	Method quantitation limit
MQO	Measurement quality objective
MRL	Method reporting limit
MSDS	Material safety data sheet
MTCA	Model Toxics Control Act
ND	Not detected
NPDES	National Pollutant Discharge Elimination System
NPL	National Priorities List
OC	Organochlorine
PA	Preliminary Assessment
PAHs	Polynuclear aromatic hydrocarbons
PARCC	Precision, accuracy, representativeness, comparability, and completeness
PCBs	Polychlorinated biphenyls
PCD	Project controlling document
PCE	Tetrachloroethene
PDM	Percent decision match
PE	Performance evaluation
PQL	Practical quantitation limit
PRGs	Preliminary remediation goals
QA	Quality assurance
QC	Quality control
QL	Quantitation limit
RA	Remedial Action
RAGS	Risk Assessment Guidance for Superfund
RBCs	Risk-based concentrations
RCRA	Resource Conservation and Recovery Act
RD	Remedial Design
Redox	Oxidation-reduction potential
RFI	RCRA Facility Investigation
RI/FS	Remedial Investigation/Feasibility Study

RL	Reporting limit
RME	Reasonable maximum exposure
RPD	Relative percent difference
RPM	Remedial project manager
RSD	Relative standard deviation
RT	Regulatory threshold
SAD	Sum of absolute deviations
SAPs	Sampling and analysis plans
SI	Site Investigation
SQL	Sample quantitation limit
SSS	Sample sum of sequences
TCE	Trichloroethene
TCLP	Toxicity characteristic leaching procedure
TIN	Triangular irregular network
TNT	Trinitrotoluene
TPH	Total petroleum hydrocarbons
TPP	Technical project planning
TSCA	Toxic Substance Control Act
UCL	Upper confidence limit
USACE	U.S. Army Corps of Engineers
UTL	Upper tolerance limit
VOCs	Volatile organic compounds
WLS	Weighted least squares

Symbols and Notations

Symbol	Description
α	Significance level of a statistical test
$\forall_{i,j}$	All *i* and *j*
b_0	Intercept estimate for linear regression
b_1	Slope estimate for linear regression
$1-\beta$	Power of a statistical test
β_0	True intercept of a regression equation
β_1	True slope of a regression equation
C	Target contaminant concentration or fixed-threshold value

Symbol	Description
CV	Coefficient of variation
e_i	Sample residual
ε	Population residual
$F_{p,k,q}$	Critical value of the *F* distribution with *k* numerator degrees of freedom and *q* denominator degrees of freedom where 100*p*% of the distribution lies below this value
γ	Population correlation coefficient
$\gamma(h)$	Semivariogram function
IQR	Sample interquartile range
H_0	Null hypothesis of a statistical test
H_A	Alternative hypothesis of a statistical test
Ln	Natural logarithm
Log	Base ten logarithm
μ	Population mean
$\hat{\mu}_1$	Minimum variance unbiased estimate (MVUE) of the population mean of a lognormal distribution
n	Number of observations in a sample
ν	Degrees of freedom (df)
p	Sample proportion or probability of an event for the binomial distribution
P	Population proportion of a random variable
$P(X)$	Probability density function of random variable *X*
$P(X_a \leq X \leq X_b)$	Probability that the random variable *X* lies between X_a and X_b
r	Pearson’s sample correlation coefficient
R	Sample range
$R(x_i)$	Rank of the i^{th} observation with respect to the other observations
ρ	Spearman’s rank order sample correlation coefficient
s	Sample standard deviation

Symbol	Description
s^2	Sample variance
σ	Population standard deviation
σ^2	Population variance
$t_{p,\nu}$	Critical value of the *t* distribution with ν degrees of freedom where 100*p*% of the distribution lies below this value
τ	Kendall’s rank order sample correlation coefficient
$\mathit{\Theta}$	A population parameter
θ	A population parameter
w_i	Number of ties in the *i*th group or *i*th weighting factor
$\bar{x}$	Sample arithmetic mean
$\tilde{x}$	Sample median
$\vec{x}_i$	A vector $\left(x_{i1}, x_{i2}, \ldots, x_{im}\right)$
$x_1, x_2, \ldots, x_n$	A set of *n* observations, a sample
$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$	A set of *n* observations ordered from least to greatest
$\chi^2_{p,\nu}$	Critical value of the chi-squared distribution with ν degrees of freedom where 100*p*% of the distribution lies below this value
x_p	100*p*th percentile or *p* quantile of a sample
X_p	100*p*th percentile or *p* quantile of random variable X
$X, Y, etc.$	Random variables representing populations
Z_p	Critical value of the standard normal distribution where 100*p*% of the distribution lies below this value