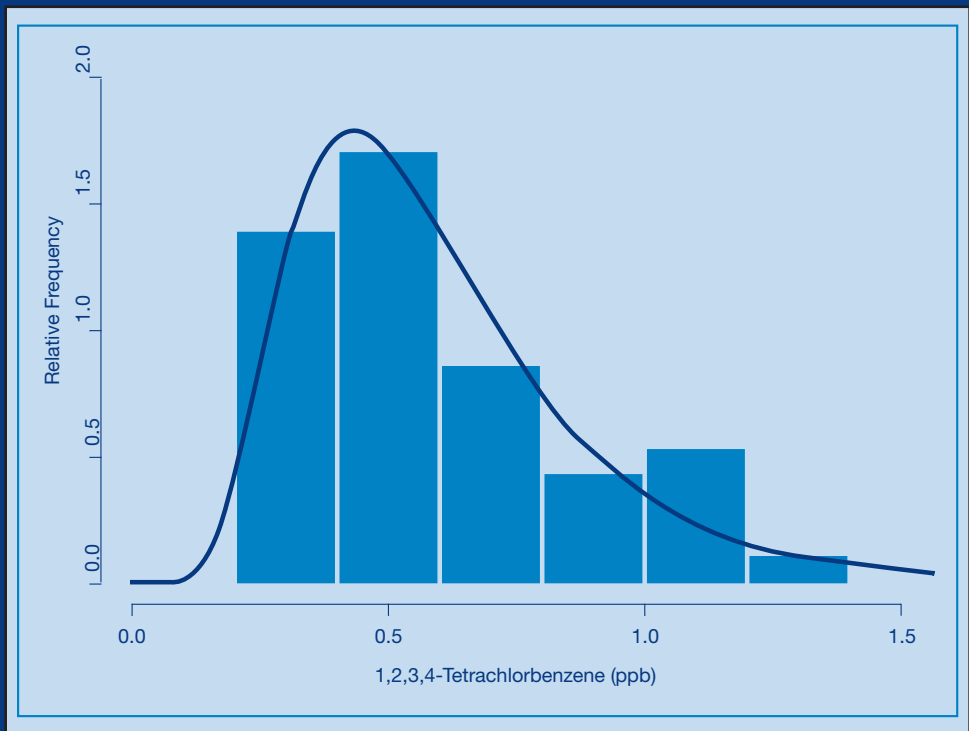


Environmental Statistics

with **S-PLUS**



***Steven P. Millard
Nagaraj K. Neerchal***

APPLIED ENVIRONMENTAL STATISTICS

Environmental Statistics

***with* S-PLUS**

APPLIED ENVIRONMENTAL STATISTICS

Steven P. Millard, *Series Editor*

Environmental Statistics with S-Plus

Steven P. Millard and Nagaraj K. Neerchal

FORTHCOMING TITLES

Groundwater Monitoring and Regulations

Charles Davis

Statistical Tools for Environmental Quality

Michael E. Ginevan and Douglas E. Splitstone

Environmental Statistics

with S-PLUS

***Steven P. Millard
Nagaraj K. Neerchal***



CRC Press

Boca Raton London New York Washington, D.C.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2000 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20141208

International Standard Book Number-13: 978-1-4200-3717-3 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

PREFACE

The environmental movement of the 1960s and 1970s resulted in the creation of several laws aimed at protecting the environment, and in the creation of Federal, state, and local government agencies charged with enforcing these laws. Most of these laws mandate monitoring or assessment of the physical environment, which means someone has to collect, analyze, and explain environmental data. Numerous excellent journal articles, guidance documents, and books have been published to explain various aspects of applying statistical methods to environmental data analysis. Only a very few books attempt to provide a comprehensive treatment of environmental statistics in general, and this book is an addition to that category.

This book is a survey of statistical methods you can use to collect and analyze environmental data. It explains *what* these methods are, *how* to use them, and *where* you can find references to them. It provides insight into what to think about *before* you collect environmental data, how to collect environmental data (via various random sampling schemes), and also how to make sense of it *after* you have it. Several data sets are used to illustrate concepts and methods, and they are available both with software and on the CRC Press Web so that the reader may reproduce the examples. The appendix includes an extensive list of references.

This book grew out of the authors' experiences as teachers, consultants, and software developers. It is intended as both a reference book for environmental scientists, engineers, and regulators who need to collect or make sense of environmental data, and as a textbook for graduate and advanced undergraduate students in an applied statistics or environmental science course. Readers should have a basic knowledge of probability and statistics, but those with more advanced training will find lots of useful information as well.

A unique and powerful feature of this book is its integration with the commercially available software package S-PLUS, a popular and versatile statistics and graphics package. S-PLUS has several add-on modules useful for environmental data analysis, including ENVIRONMENTALSTATS for S-PLUS, S+SPATIALSTATS, and S-PLUS for ArcView GIS. Throughout this book, when a data set is used to explain a statistical method, the commands for and results from the software are provided. Using the software in conjunction with this text will increase the understanding and immediacy of the methods.

This book follows a more or less sequential progression from elementary ideas about sampling and looking at data to more advanced methods of estimation and testing as applied to environmental data. Chapter 1 provides an introduction and overview, Chapter 2 reviews the Data Quality Objectives (DQO) and Data Quality Assessment (DQA) process necessary in the design

and implementation of any environmental study, and Chapter 3 talks about exploratory data analysis (EDA). Chapter 4 explains the idea of a population, sample, random variable, and probability distribution. Chapter 5 details various methods for estimating characteristics of a population (probability distribution) based on a sample (data). Chapter 6 discusses prediction intervals, tolerance intervals, and control charts, which have been used in the manufacturing industry for a long time and have been proposed as good methods to use in groundwater monitoring. Chapter 7 reviews the basic ideas in hypothesis testing, including balancing the two possible errors a decision maker can make (e.g., declaring a site contaminated when it really is not, or declaring a site not contaminated when it really is). This chapter also illustrates tests for goodness-of-fit and outliers, classical and nonparametric methods for comparing one, two, or several groups (e.g., background vs. potentially contaminated sites), and the multiple comparisons problem. Chapter 8 returns to the DQO process of Chapter 2 and illustrates how to determine required sample sizes based on the statistical theory presented in Chapters 6 and 7. Chapter 9 discusses linear models, including correlation, simple regression, testing for trend, and multiple regression. This chapter also explains the idea of calibration and how this relates to measuring chemical concentrations and determining various limits associated with the chemical measurement process (i.e., decision limit, detection limit, and quantitation limit). Chapter 10 continues the ideas on calibration discussed in Chapter 9 by explaining how to handle environmental data that contain “less-than-detection-limit” results. Chapter 11 examines methods for dealing with data collected over time that may be serially correlated. Chapter 12 considers how to handle data collected over space that may be spatially correlated. Finally, Chapter 13 discusses the immense field of risk assessment, which usually involves both “hard” data and expert judgment.

TYPOGRAPHIC CONVENTIONS

Throughout this book, we use the following typographic conventions:

- The **bold font** is used for section headings, figure and table titles, equation numbers, and what you click on dialog boxes.
- The *italic font* and ***bold italic font*** are used for emphasis.
- The `courier font` is used to display commands that you type into the S-PLUS Command or Script Window, and the names of variables and functions within S-PLUS (S-PLUS objects).
- The *italic courier font* is used for mathematical notation (e.g., variable names, function definitions, etc.).

A NOTE ABOUT S-PLUS AND GRAPHICS

Throughout this book we assume the reader has access to S-PLUS and ENVIRONMENTALSTATS for S-PLUS, knows how to start S-PLUS, and knows how to load the ENVIRONMENTALSTATS for S-PLUS module. Also, in Chapter 12 where we deal with spatial statistics, we assume the reader has access to and is running S+SPATIALSTATS, ArcView GIS, and S-PLUS for ArcView GIS.

At the time this book was being written, the current version of S-PLUS for Windows was S-PLUS 2000 Release 2, and the current version of S-PLUS for UNIX was Version 5.1. The current version of ENVIRONMENTALSTATS for S-PLUS was Version 1.1 Release 2, but Version 2.0 (which includes pull-down menus for the Windows version of S-PLUS) was in Beta Release and should be available by the time this book is published. Throughout this book, we have included examples demonstrating how to use S-PLUS and ENVIRONMENTALSTATS for S-PLUS to create the figures and analyses shown in this book. We assume the reader is using one of the above-mentioned versions of S-PLUS or a later version.

The current UNIX version of S-PLUS only works at the command line, so if you use this version of S-PLUS you can safely ignore sections that begin with the heading Menu (the next UNIX version of S-PLUS, however, will include pull-down menus). If you use the Standard Edition of S-PLUS for Windows, you can only use the pull-down menus and toolbars; you do not have access to the command line, so you can safely ignore sections that begin with the heading Command. If you use the Professional Edition of S-PLUS for Windows, you can use both the pull-down menus and toolbars and the command line, so you can apply the information listed under both headings.

Many of the examples of using the command line under a Command heading use the `attach` function to attach a data frame to your search list. This is done in order to be able to reference the columns of the data frame explicitly without having to use subscript operators. Please be aware that it is possible you may have a data object in your working directory with the same name as the column of the data frame that is being used, in which case your data object will “mask” the column of the data frame. For example, the data frame `epa.94b.tccb.df` contains a column named `Area`. If you already have a data object named `Area` in your working directory, then any examples that involve attaching `epa.94b.tccb.df` and using the column `Area` will not work correctly. In these cases, you must change the name of your data object or use the `$` or `[` operator to direct S-PLUS to the correct data set (e.g., `epa.94b.tccb.df$Area`).

In S-PLUS and ENVIRONMENTALSTATS for S-PLUS it is very easy to produce color plots. In fact, most of the built-in plotting functions produce

color plots by default. In this book, however, all of the plots are black and white or grayscale due to the high cost of color printing. The steps for producing color plots are still included in the examples in this book, but the pictures in the book will be in black and white, whereas in many cases the pictures on your computer screen will be in color.

All of the graphs you can create with ENVIRONMENTALSTATS for S-PLUS use traditional S-PLUS graphics and, as such, they are not editable in the Windows version of S-PLUS. If you use the Windows version of S-PLUS, you can convert traditional plots to editable plots by right-clicking on the data part of the graph and choosing Convert to Objects from the context menu.

ABOUT THE AUTHORS

Steven P. Millard was born in Williamsburg, Virginia and raised in Arlington, Virginia. He received a bachelor's degree in mathematics from Pomona College in Claremont, California in 1980, and a Ph.D. in biostatistics from the University of Washington, Seattle in 1995. He has taught at the University of California at Santa Barbara and Saint Martin's College in Lacey, Washington. He ran a consulting unit at the University of Washington and also worked for CH2M Hill on the second Love Canal Habitability Study. From 1990 to 1993 Dr. Millard ran the training program in S-PLUS at Statistical Sciences, Inc. (now part of MathSoft, Inc.). From 1993 through the end of the century, he was a private statistical consultant through his company Probability, Statistics & Information (PSI), working on projects ranging from analyzing water quality data from the Everglades National Park to developing software for automated housing appraisal. Currently, Dr. Millard is the Manager of Consulting Services for the Data Analysis Products Division of MathSoft, Inc. He is the creator of ENVIRONMENTALSTATS for S-PLUS, an add-on module to S-PLUS for environmental data analysis.

Nagaraj K. Neerchal was born in a west coast village in southern India. He received his bachelor's and master's degrees in statistics, in 1981 and 1982, respectively, from the Indian Statistical Institute in Calcutta. He received a Ph.D. in statistics from Iowa State University, Ames in 1986. That same year he joined the Department of Mathematics and Statistics at the University of Maryland Baltimore County, where he is currently a Professor of Statistics. He served as the interim chair of the department 1999 – 2000. Dr. Neerchal's main areas of research interest are time series analysis and methods of analyzing correlated categorical data. He has worked with various environmental applications involving pollution monitoring, sampling, and data quality issues related to policy making. Dr. Neerchal has done extensive consulting for various government agencies and private organizations on statistical problems. He was a Senior Research Scientist at Pacific Northwest National Laboratories, Richland, Washington.

ACKNOWLEDGMENTS

This book is the result of knowledge we have acquired over the course of our careers. Several people were instrumental in imparting this knowledge to us. Steve Millard would like to thank Don Bentley (Mathematics Department, Pomona College), Dennis Lettenmaier (Department of Civil Engineering, University of Washington), and Peter Guttorp (Department of Statistics, University of Washington) for their guidance in his statistical education, and all of the people at MathSoft and on S-news who have responded to his many questions about S-PLUS. Nagaraj Neerchal gratefully acknowledges the continued support of Drs. Phil Ross, Barry Nussbaum, Ron Shafer, and Pepi Lacayo and the constant encouragement of Professor Bimal Sinha.

Several people have helped us prepare this book. We thank Robert Stern at Chapman & Hall/CRC Press for all of his time and help as the managing editor, and Chris Andreasen, our copy editor, for her careful review of the manuscript. We are also grateful to Dr. Richard O. Gilbert of Battelle for his reviews and suggestions, and his help in general as a colleague. We are grateful to Charles Davis, Henry Kahn, and Bruce Peterson for reviewing drafts of some of the chapters. A cooperative agreement between the University of Maryland Baltimore County and the U.S. Environmental Protection Agency supported some of Nagaraj Neerchal's work that was incorporated into this book.

Data used in the examples of Chapters 11 and 12 were from Nagaraj Neerchal's collaborations with Dr. Susan Brunenmeister of the Chesapeake Bay Program and Dr. Steve Weisberg of Versar Inc. (now with the Southern California Water Research Project). We thank them for their collaboration.

Finally, we would like to thank our families, Stacy Selke and Chris Millard, and Chetana, Harsha, and Siri Neerchal, for their continued love and support of our endeavors.

TABLE OF CONTENTS

1	Introduction	1
	Intended Audience.....	2
	Environmental Science, Regulations, and Statistics.....	2
	Overview	7
	Data Sets and Case Studies	10
	Software	11
	Summary	12
	Exercises	12
2	Designing a Sampling Program, Part I.....	13
	The Basic Scientific Method	13
	What is a Population and What is a Sample?	15
	Random vs. Judgment Sampling	15
	The Hypothesis Testing Framework	16
	Common Mistakes in Environmental Studies	17
	The Data Quality Objectives Process	19
	Sources of Variability and Independence.....	24
	Methods of Random Sampling.....	26
	Case Study.....	42
	Summary	49
	Exercises	51
3	Looking at Data	53
	Summary Statistics	53
	Graphs for a Single Variable	67
	Graphs for Two or More Variables	113
	Summary	133
	Exercises	134
4	Probability Distributions	139
	What is a Random Variable?.....	139
	Discrete vs. Continuous Random Variable.....	140
	What is a Probability Distribution?	141
	Probability Density Function (PDF).....	145
	Cumulative Distribution Function (CDF).....	153
	Quantiles and Percentiles	158
	Generating Random Numbers from Probability Distributions	161
	Characteristics of Probability Distributions	162
	Important Distributions in Environmental Statistics	167
	Multivariate Probability Distributions.....	194
	Summary	194
	Exercises	195

5	Estimating Distribution Parameters and Quantiles	201
	Methods for Estimating Distribution Parameters	201
	Using ENVIRONMENTALSTATS for S-PLUS to Estimate Distribution Parameters	215
	Comparing Different Estimators	219
	Accuracy, Bias, Mean Square Error, Precision, Random Error, Systematic Error, and Variability	225
	Parametric Confidence Intervals for Distribution Parameters	228
	Nonparametric Confidence Intervals Based on Bootstrapping.....	257
	Estimates and Confidence Intervals for Distribution Quantiles (Percentiles)	274
	A Cautionary Note about Confidence Intervals.....	289
	Summary	290
	Exercises	292
6	Prediction Intervals, Tolerance Intervals, and Control Charts	295
	Prediction Intervals	296
	Simultaneous Prediction Intervals	320
	Tolerance Intervals	335
	Control Charts	353
	Summary	360
	Exercises	361
7	Hypothesis Tests	365
	The Hypothesis Testing Framework	365
	Overview of Univariate Hypothesis Tests.....	371
	Goodness-of-Fit Tests	371
	Test of a Single Proportion.....	385
	Tests of Location	389
	Tests on Percentiles	409
	Tests on Variability	410
	Comparing Locations between Two Groups: The Special Case of Paired Differences	412
	Comparing Locations between Two Groups	415
	Comparing Two Proportions	441
	Comparing Variances between Two Groups.....	446
	The Multiple Comparisons Problem	450
	Comparing Locations between Several Groups	453
	Comparing Proportions between Several Groups	461
	Comparing Variability between Several Groups	462
	Summary	466
	Exercises	467
8	Designing a Sampling Program, Part II	471
	Designs Based on Confidence Intervals	471

Designs Based on Nonparametric Confidence, Prediction, and Tolerance Intervals	481
Designs Based on Hypothesis Tests	485
Optimizing a Design Based on Cost Considerations	521
Summary	523
Exercises	524
9 Linear Models	527
Covariance and Correlation	527
Simple Linear Regression	539
Regression Diagnostics	553
Calibration, Inverse Regression, and Detection Limits	562
Multiple Regression	575
Dose-Response Models: Regression for Binary Outcomes	584
Other Topics in Regression	588
Summary	589
Exercises	590
10 Censored Data	593
Classification of Censored Data	593
Graphical Assessment of Censored Data	597
Estimating Distribution Parameters	609
Estimating Distribution Quantiles	636
Prediction and Tolerance Intervals	637
Hypothesis Tests	640
A Note about Zero-Modified Distributions	645
Summary	645
Exercises	645
11 Time Series Analysis	647
Creating and Plotting Time Series Data	647
Autocorrelation	651
Dealing with Autocorrelation	669
More Complicated Models:	
Autoregressive and Moving Average Processes	671
Estimating and Testing for Trend	672
Summary	689
Exercises	690
12 Spatial Statistics	693
Overview: Types of Spatial Data	693
The Benthic Data	694
Models for Geostatistical Data	700
Modeling Spatial Correlation	703
Prediction for Geostatistical Data	721

Using S-PLUS for ArcView GIS	727
Summary	733
Exercises	734
13 Monte Carlo Simulation and Risk Assessment.....	735
Overview	736
Monte Carlo Simulation	736
Generating Random Numbers	741
Uncertainty and Sensitivity Analysis	748
Risk Assessment.....	758
Summary	776
Exercises	777
References	779

1 INTRODUCTION

The environmental movement of the 1960s and 1970s resulted in the creation of several laws aimed at protecting the environment, and in the creation of Federal, state, and local government agencies charged with enforcing these laws. In the U.S., laws such as the Clean Air Act, the Clean Water Act, the Resource Conservation and Recovery Act, and the Comprehensive Emergency Response and Civil Liability Act mandate some sort of monitoring or comparison to ensure the integrity of the environment. Once you start talking about monitoring a process over time, or comparing observations from two or more sites, you have entered the world of numbers and statistics. In fact, more and more environmental regulations are mandating the use of statistical techniques, and several excellent books, guidance documents, and journal articles have been published to explain how to apply various statistical methods to environmental data analysis (e.g., Berthouex and Brown, 1994; Gibbons, 1994; Gilbert, 1987; Helsel and Hirsch, 1992; McBean and Rovers, 1998; Ott, 1995; Piegorsch and Bailer, 1997; ASTM, 1996; USEPA, 1989a,b,c; 1990; 1991a,b,c; 1992a,b,c,d; 1994a,b,c; 1995a,b,c; 1996a,b; 1997a,b). Only a very few books attempt to provide a comprehensive treatment of environmental statistics in general, and even these omit some important topics.

This explosion of regulations and mandated statistical analysis has resulted in at least four major problems.

- Mandated procedures or those suggested in guidance documents are not always appropriate, or may be misused (e.g., Millard, 1987a; Davis, 1994; Gibbons, 1994).
- Statistical methods developed in other fields of research need to be adapted to environmental data analysis, and there is a need for innovative methods in environmental data analysis.
- The backgrounds of people who need to analyze environmental data vary widely, from someone who took a statistics course decades ago to someone with a Ph.D. doing high-level research.
- There is no single software package with a comprehensive treatment of environmental statistics.

This book is an attempt to solve some of these problems. It is a survey of statistical methods you can use to collect and analyze environmental data. It explains *what* these methods are, *how* to use them, and *where* you can find references to them. It provides insight into what to think about *before* you collect environmental data, how to collect environmental data (via various

random sampling schemes), and also and how to make sense of it *after* you have it. Several data sets are used to illustrate concepts and methods, and are available in ENVIRONMENTALSTATS for S-PLUS (see below) and/or on the CRC Press Web site at www.crcpress.com so that the reader may reproduce the examples. You will also find a list of relevant URLs on this Web site. The appendices of this book include an extensive list of references and an index.

A unique and powerful feature of this book is its integration with the commercially available software package S-PLUS, a popular and powerful statistics and graphics package. S-PLUS has several add-on modules useful for environmental data analysis, including ENVIRONMENTALSTATS for S-PLUS, S+SPATIALSTATS, and S-PLUS for ArcView GIS. Throughout this book, when a data set is used to explain a statistical method, the commands for and results from the software are provided. Using the software in conjunction with this text will increase the understanding and immediacy of the methods.

INTENDED AUDIENCE

This book grew out of the authors' experience as teachers, consultants, and software developers. It is intended as both a reference book for environmental scientists, engineers, and regulators who need to collect or make sense of environmental data, and as a textbook for graduate and advanced undergraduate students in an applied statistics or environmental science course. Readers should have a basic knowledge of probability and statistics, but those with more advanced training will find lots of useful information as well. Readers will find that topics are introduced at an elementary level, but the theory behind the methods is explained as well, and all pertinent equations are included. Each topic is illustrated with examples.

ENVIRONMENTAL SCIENCE, REGULATIONS, AND STATISTICS

As a brief introduction to some of the problems involved in environmental statistics, this section discusses three examples where environmental science, regulations, and statistics intersect. Each of these examples illustrates several issues that need to be considered in sampling design and statistical analysis. We will discuss many of these issues both in general terms and in detail throughout this book.

Groundwater Monitoring at Hazardous and Solid Waste Sites

The Resource Conservation and Recovery Act (RCRA) requires that groundwater near hazardous waste sites and municipal solid waste sites be

monitored to ensure that chemicals from the site are not leaking into the groundwater (40 CFR Parts 264 and 265; 40 CFR Part 258). So how do you design a program to monitor groundwater?

Several Federal and state guidance documents have been published addressing the design and statistical issues associated with complying with RCRA regulations (e.g., USEPA, 1989b; 1991c; 1992b,c). The current practice at most sites is to start in a phase called **detection monitoring** in which groundwater is sampled at **upgradient** and **downgradient** wells a number of times each year. The groundwater at the upgradient wells is supposed to represent the quality of **background** groundwater that has not been affected by leakage from the site. Figure 1.1 is a simple schematic for the physical setup of such a monitoring program. More detailed figures can be found in Sara (1994, Chapters 9 to 11) and Gibbons (1995, p. 186).

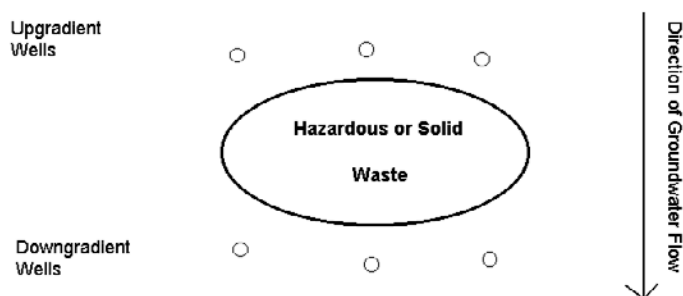


Figure 1.1 Simple schematic of an aerial view of a groundwater monitoring system

During detection monitoring, the groundwater samples are analyzed for **indicator parameters** such as pH, conductance, total organic carbon, and total organic halides. For each indicator parameter and each downgradient well, the value of the parameter at the downgradient well is compared to the value at the upgradient well(s). If the value of the parameter is deemed to be “above background” then, depending on the permit, the owner/operator of the site may be required to take more samples, or the site may enter a second phase of monitoring called **assessment monitoring** or **compliance monitoring**.

In assessment or compliance monitoring, the owner/operator of the site is required to start analyzing the groundwater samples for other chemical parameters, such as the concentrations of specific chemicals. The concentrations of chemicals in the groundwater from downgradient wells are compared to fixed concentration limits (Ground Water Protection Standards or GWPS) such as a Maximum Contaminant Level (MCL) or Alternative Concentration Limit (ACL). For any specified chemical, if the concentration is “above” the GWPS, the site enters a third phase called *corrective action monitoring*.

There are several basic scientific design and statistical issues involved in monitoring groundwater, including:

- How do you determine what constitutes an upgradient well and what constitutes a downgradient well? Can you be sure the gradient will stay the same over time?
- What chemicals are contained in the site? Are the mandated indicator parameters for detection monitoring good indicators of leakage of these particular chemicals?
- During assessment monitoring, are you required to test for chemicals that are not contained in the site? If so, why?
- For detection monitoring, how do you determine whether an indicator parameter at a downgradient well is “above” background? For each indicator parameter, what kind of increase in value is important to detect and how soon?
- Is there “significant” spatial and/or temporal variability in any of the indicator parameters or chemical concentrations in the upgradient area? If so, how do you account for this when comparing upgradient and downgradient wells? Is it possible to use intrawell comparisons instead of comparing downgradient wells with upgradient wells?
- What other sources of random variation are present? Is there a lot of variability between samples taken on the same day? Is there a lot of variability in the chemical measurement process? Are different laboratories being used, and if so, is there a lot of variability between labs?
- For assessment monitoring, what is the basis of each GWPS? How do you tell whether chemical concentrations at downgradient wells are “above” the GWPS?
- How do you account for the possibility of false alarms, which involve increased monitoring costs, and the possibility of missing contamination, which involves a potential threat to public health?

Soil Cleanup at Superfund Sites

The Comprehensive Emergency Response and Civil Liability Act (CERCLA), also known as “Superfund,” requires a remedial investigation/feasibility study (RI/FS) at each site on the National Priorities List (NPL) to determine the extent of contamination and the risks posed to human health and the environment. The U.S. Environmental Protection Agency (USEPA) has developed several guidance documents that discuss design and analysis issues for various stages of this process (USEPA, 1987a,b; 1989a,c; 1991b; 1992b,d; 1994b; 1996b,c).

The guidance documents **Soil Screening Guidance: User’s Guide** (USEPA, 1996b) and **Soil Screening Guidance: Technical Background Document** (USEPA, 1996c) discuss the use of Soil Screening Levels (SSLs) at Superfund sites that may have future residential land use to determine whether soil in a particular area requires further investigation and/or remediation, or if it can be left alone (or at least does not require any further attention under CERCLA). The guidance suggests stratifying the site into areas that are contaminated, areas unlikely to be contaminated, and areas that may be contaminated. Within each stratum, the guidance suggests dividing the area into exposure areas (EAs) that are up to a half acre in size, and taking soil samples within each EA. For each EA, the concentration of a particular chemical of concern is compared with the SSL. If the concentration is “greater” than the SSL, then the EA requires further investigation.

This soil screening guidance involves several basic scientific design and statistical issues, including:

- How do you know what chemicals you are looking for?
- How do you know the boundary of the area to look at? How do you determine which areas are contaminated, which are not, and which might be?
- What is the basis for an SSL for a particular chemical?
- How do you determine whether the chemical concentration in the soil is “greater” than the SSL?
- What are the sources of random variation in the data (e.g., field variability, collector variability, within lab variability, between lab variability, etc.), and how do you account for them?
- How do you account for the possibility of false alarms (saying the chemical concentration is greater than the SSL when it is not), which involves unnecessary costs for further investigation, and the possibility of missing contamination (saying the chemical concentration is less than the SSL when in fact it not), which involves a potential threat to public health?

Monitoring Air Quality

The Clean Air Act is the comprehensive Federal law that regulates air emissions from area, stationary, and mobile sources. This law authorizes the USEPA to establish National Ambient Air Quality Standards (NAAQS), which are national targets for acceptable concentrations of specific pollutants in the air.

There are two kinds of standards: primary and secondary. Primary standards set limits to protect public health, including the health of “sensitive” populations such as asthmatics, children, and the elderly. Secondary standards set limits to protect public welfare, including effects on soils, water, crops, vegetation, buildings, property, animals, wildlife, weather, visibility, transportation, and other economic values, as well as personal comfort and well-being.

The USEPA has set national air quality standards for seven principal air pollutants, called *criteria air pollutants*: carbon monoxide (CO), lead (Pb), nitrogen dioxide (NO₂), volatile organic compounds (VOCs), ozone (O₃), particulate matter (PM-10), and sulfur dioxide (SO₂). As an example, the primary standard for ozone is based on comparing a specific limit to the daily maximum ozone measurement among the network of stations monitoring a specific area. The monitors at each station record ozone concentrations continuously in time, so there are several possible ways to define the “daily maximum ozone concentration.” Between 1978 and 1997, the daily maximum ozone concentration was based on averaging the concentrations for each hour to produce 24 observations per station per day, and the maximum daily concentration at a station was defined to be the maximum of these 24 values. The daily maximum value (over all of the monitoring stations) could exceed 0.12 parts per million (ppm) only three times or fewer within a 3-year period.

The proposed new standard for ozone divides the 24-hour day into three 8-hour blocks, concentrations are averaged within each 8-hour block to produce three observations per station per day, and the maximum daily concentration at a station is defined to be the maximum of these three values. The standard looks at the fourth highest daily maximum concentration over all stations within a single year. If the 3-year average of these annual fourth highest daily maximum concentrations is less than 0.08 ppm, then the standard is met.

The scientific basis for national ambient air quality standards depends on knowledge about the effects of air pollutants on health and the environment, which involves clinical, epidemiological, and field studies, all of which depend on scientific design and statistical analysis. Also, USEPA monitors trends in air quality over time, which involves time series analysis. Nychka et al. (1998) present several different types of statistical analyses of air quality data.

OVERVIEW

This section explains the background and layout of this book.

What is Environmental Statistics?

Environmental statistics is simply the application of statistical methods to problems concerning the environment. Examples of activities that require the use of environmental statistics include:

- Monitoring air or water quality.
- Monitoring groundwater quality near a hazardous or solid waste site.
- Using risk assessment to determine whether an area with potentially contaminated soil needs to be cleaned up, and, if so, how much.
- Assessing whether a previously contaminated area has been cleaned up according to some specified criterion.
- Using hydrological data to predict the occurrences of floods.

The term “environmental statistics” must also include work done in various branches of ecology, such as animal population dynamics and general ecological modeling, as well as other fields, such as geology, chemistry, epidemiology, oceanography, and atmospheric modeling. This book concentrates on statistical methods to analyze chemical concentrations and physical parameters, usually in the context of mandated environmental monitoring.

Environmental statistics is a special field of statistics. Probability and statistics deal with situations in which the outcome is not certain. They are built upon the concepts of a *population* and a *sample* from the population.

Probability deals with predicting the characteristics of the sample, given that you know the characteristics of the population (e.g., what is the probability of picking an ace out of a deck of 52 well-shuffled standard playing cards?). *Statistics* deals with inferring the characteristics of the population, given information from one or more samples from the population (e.g., after 100 times of randomly choosing a card from a deck of 20 unknown playing cards and then replacing the card in the deck, no ace has appeared; therefore the deck probably does not contain any aces).

The field of environmental statistics is relatively young and employs several statistical methods that have been developed in other fields of statistics, such as sampling design, exploratory data analysis, basic estimation and hypothesis testing, quality control, multiple comparisons, survival analysis, and Monte Carlo simulation. Nonetheless, special problems have motivated innovative research, and both traditional and new journals now report on statistical methods that have been developed in the context of environmental monitoring. (See Appendix A: References for a comprehensive list of journal articles, guidance documents, and general textbooks that deal with environmental statistics and related topics.)

Where Do the Data Come from?

If a law says that you have to monitor the environment because your factory is emitting chemicals into the air or water, or because you run a hazardous waste site and you have to make sure nothing is seeping into the groundwater, or because you want to develop a shopping mall on a piece of real estate that used to be occupied by a plant that put creosote on railroad ties, then you have to figure out how to collect and analyze the data. But before you collect the data, you have to figure out what the question is that you are trying to answer.

The *Data Quality Objectives (DQO) Process* is a formal method for deciding what the question is, what data need to be collected to answer the question, how the data will be collected, and how a decision will be made based on the data. It is the first and most important step of any environmental study. The DQO process is discussed in Chapter 2, along with various methods of random sampling.

The actual collection and laboratory analysis of a physical sample from the environment that eventually leads to a reported number (e.g., concentration of trichloroethylene) involves several steps (Clark and Whitfield, 1994; Ward et al., 1990, p. 19):

1. Sample collection
2. Sample handling
3. Transportation
4. Sample receipt and storage at laboratory
5. Sample work up
6. Sample analysis
7. Data entry
8. Data manipulation
9. Data reporting

Figure 1.2, reproduced from Clark and Whitfield (1994, p. 1065), illustrates these steps, along with sources of variability that cause the measured concentration to deviate from the true concentration. Sources of variability are discussed in more detail in Chapter 2.

Once the physical samples are collected, they must be analyzed in the laboratory to produce measures of chemical concentrations and/or physical parameters. This is a whole topic in itself. An environmental chemist does not just scoop a piece of dirt out of a collection vial, dissolve it in water or chemicals in a test tube, stick the tube in a machine, and record the concentrations of all the chemicals that are present. In fact, the process of measuring chemical concentrations in soil, water, or air has its own set of DQO steps! Chapter 9 discusses some of the aspects of chemometrics and the important topic of machine calibration and detection limit.

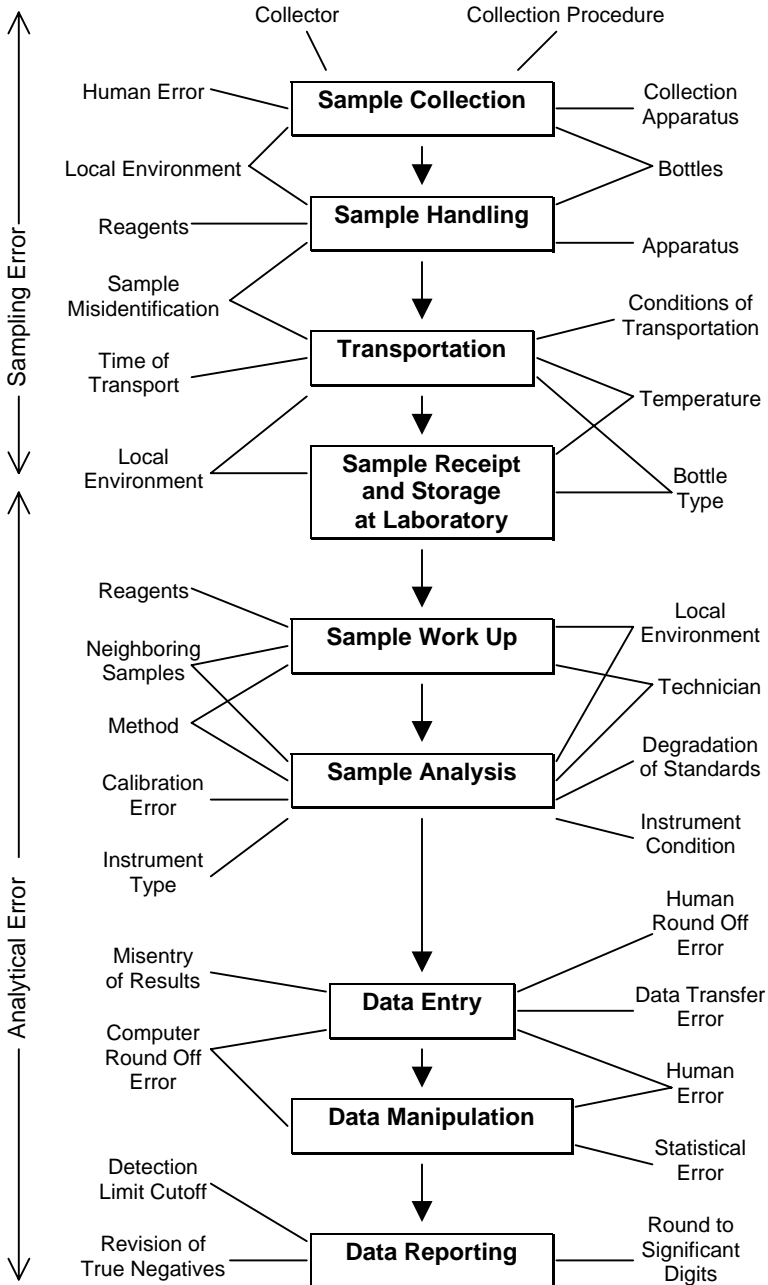


Figure 1.2

The steps involved in producing environmental data, and their associated sources of variability. (From Clark and Whitfield, 1994. *Water Resources Bulletin* 30(6), 1063–1079. With permission of American Water Resources Association, Herndon, VA.)

How Do You Analyze the Data Once You Have It?

When most people think of environmental statistics, they think of the data analysis part. The design of the sampling program and the DQO process are much more important, however. In fact, one of the steps in the DQO process is to specify how you will analyze the data once you have it.

Chapters 3 to 9 discuss various ways of analyzing environmental data, starting with exploratory data analysis (EDA) in Chapter 3. Chapter 4 explains the idea of a population, sample, random variable, and probability distribution, and gives examples of each of these in the context of a real data set. Chapter 5 discusses various methods for estimating characteristics of a population (probability distribution) based on a sample (data). Chapter 6 discusses prediction intervals, tolerance intervals, and control charts, which have been proposed as good methods to use in groundwater monitoring. Chapter 7 reviews the basic ideas in hypothesis testing, including balancing the two possible errors a decision-maker can make (e.g., declaring a site contaminated when it really is not, or declaring a site not contaminated when it really is). This chapter also illustrates tests for goodness-of-fit and outliers, classical and nonparametric methods for comparing one, two, or several groups (e.g., background vs. potentially contaminated sites), and the multiple comparisons problem. Chapter 8 returns to the DQO process of Chapter 2 and illustrates how to determine required sample sizes based on the statistical theory presented in Chapters 6 and 7. Chapter 9 discusses linear models, including correlation, simple regression, testing for trend, and multiple regression. In addition, this chapter explains the idea of calibration and how this relates to measuring chemical concentrations and determining various limits associated with the process (i.e., decision limit, detection limit, quantitation limit, etc.).

Special Topics

Chapter 10 continues the ideas on calibration discussed in Chapter 9 by explaining how to handle environmental data that contain “less-than-detection-limit” results. Chapter 11 examines methods for dealing with data collected over time that may be serially correlated. Chapter 12 considers how to handle data collected over space that may be spatially correlated. Finally, Chapter 13 discusses the immense field of risk assessment, which usually involves both “hard” data and expert judgment.

DATA SETS AND CASE STUDIES

Throughout this book, we use several data sets to illustrate statistical concepts and methods of analyzing environmental data. Many of these data sets are taken from regulatory guidance documents, but a few of them are larger data sets from the real world of environmental monitoring.

SOFTWARE

As mentioned earlier, throughout this book we use the software package S-PLUS and some of its add-on modules to display computations and graphs. S-PLUS is a popular and powerful software program for statistical and graphical analysis. You can produce results either by using drop-down menus and toolbars, or by typing commands in a command window. ENVIRONMENTALSTATS for S-PLUS is an add-on module to S-PLUS that provides several graphical and statistical methods that are used extensively in environmental statistics. S+SPATIALSTATS is an add-on module for statistical analysis of spatial data and includes kriging. S-PLUS for ArcView GIS is an add-on module that lets you link the statistical and graphical tools in S-PLUS with the mapping and visual tools of ArcView. In this book, when a data set is used to explain a statistical method, the commands for and results from the software are provided. Information on the software providers is listed below.

- **S-PLUS, ENVIRONMENTALSTATS FOR S-PLUS, S+SPATIALSTATS, and S-PLUS for ArcView GIS**

MathSoft, Inc.
Data Analysis Products Division
1700 Westlake Ave N, Suite 500
Seattle, WA 98109 U.S.A.
800-569-0123
mktg@splus.mathsoft.com
www.splus.mathsoft.com

MathSoft International
Knightway House, Park Street
Bagshot, Surrey GU19 5AQ
United Kingdom
+44 1276 475350
splus@mathsoft.co.uk
www.splus.mathsoft.com

- **ArcView GIS**

Environmental Systems Research Institute, Inc. (ESRI)
380 New York Street
Redlands, CA 92373-8100 U.S.A.
909-793-2853
www.esri.com

SUMMARY

- Several laws mandate some sort of monitoring or comparison to ensure the integrity of the environment.
- This explosion of regulations and mandated statistical analysis has resulted in several problems, including inappropriate or misused statistical procedures, a need for more research, a wide variety of backgrounds in the people who need to use environmental statistics, and a lack of comprehensive software. This book is an attempt to address these problems.
- Probability deals with predicting the characteristics of the sample, given that you know the characteristics of the population.
- Statistics deals with inferring the characteristics of the population, given information from one or more samples from the population.
- Environmental statistics is the application of statistical methods to problems concerning the environment, such as monitoring air or water quality, or comparing chemical concentrations at two sites.
- This book discusses the importance of determining the question that needs to be answered, planning the design of an environmental study, how to look at data once you have collected it, and how to use statistical methods to help in the decision-making process.
- S-PLUS and its add-on modules, including ENVIRONMENTALSTATS for S-PLUS, are useful tools for the analysis of environmental data.

EXERCISES

- 1.1. Compile a partial list of Federal, state, and local agencies that deal with the environment. Two good places to start are the government section of the telephone book, and national agency sites on the World Wide Web. (The URL of the U.S. Environmental Protection Agency is www.epa.gov. Also, www.probstatinfo.com provides links to agencies that deal with the environment.)
- 1.2. Compile a partial list of national, state, and local regulations that require some sort of environmental monitoring.
- 1.3. Look in the Yellow Pages (published or on the Web) under the heading “Environmental.” What kinds of listings are there?

2 DESIGNING A SAMPLING PROGRAM, PART I

The DQO Process

The first and most important step of any environmental study is to define the objectives of the study and to design the sampling program. This chapter discusses the basic scientific method and the Data Quality Objectives (DQO) process, which is a formal mechanism for implementing the scientific method and identifying important information that needs to be known in order to make a decision based on the outcome of the study (e.g., clean up the site or leave it alone). One of the major steps in the DQO process involves deciding how you will sample the environment in order to produce the information you need to make a decision. We therefore also discuss several sampling methods in this chapter as well. Finally, we present a real-world case study to illustrate the DQO process.

THE BASIC SCIENTIFIC METHOD

Any scientific study, whether it involves monitoring pollutants in the environment, determining the efficacy of a new drug, or attempting to improve the precision of airplane parts, can eventually be boiled down to trying to determine a cause and effect relationship. Over the centuries, science has developed a set of rules to follow to try to rationally determine cause and effect. These rules can be summarized as follows:

1. Form a hypothesis about the relationship between the supposed cause and the observed effect (e.g., the presence of a pollutant in a river has decreased the population of a particular species of fish).
2. Perform an experiment in which one set of subjects (e.g., fish, people, petri dishes, etc.) is exposed to the cause, and another set of subjects experiences exactly the same conditions as the first set of subjects, except they are not exposed to the cause. The group of subjects exposed to the cause is termed the ***experimental group*** or ***exposed group***, and the other group is termed the ***control group***. All subjects must be similar to one another and be randomly assigned to the experimental or control group.
3. Record and analyze the results of the experiment.
4. Revise the hypothesis based on the results. Repeat Steps 2 to 4.

The scientific method recognizes the fact that our environment is constantly changing and that any of these changes may create an observed effect which may or may not be related to some cause we have hypothesized. The only way to determine whether a specific cause creates a specific effect is through careful experimentation that matches the experimental group with the control group on every possible condition except for allowing the experimental group to be exposed to the cause. In reality, this is often extremely difficult or impossible to achieve.

Observational Experiments

Often in environmental studies, the experiment is not actually controlled by scientists doing the investigation, but rather it is an ***observational experiment*** (often called an ***epidemiological study***) in which the experimental group is a group of people, organisms, aquifers, etc. that has been exposed to a cause (e.g., a pollutant) by virtue of physical location or other factors. A control group is then selected based upon the characteristics of the experimental group. An observational experiment can be very useful for initially identifying possible causes and effects, but suffers from the major drawback that the experimental group is self-selected, rather than randomly assigned, and therefore there might be something peculiar about this group that caused the observed effect, rather than the hypothesized cause. The tobacco industry used this argument for years to claim that there is no “proof” that smoking causes cancer, since all studies on smoking and cancer in humans are observational experiments (smoking was not randomly assigned to one group and no smoking to another). On the other hand, if many, many observational experiments result in the same conclusions, then this is very good evidence of a direct cause and effect.

The Necessity of a Good Sampling Design

No amount of advanced, cutting-edge statistical theory and techniques can rescue a study that has produced poor quality data, not enough data, or data irrelevant to the question it was meant to answer. From the very start of an environmental study, there must be a constant dialog between the data producers (field and lab personnel, data coders, etc.), the data users (scientists and statisticians), and the ultimate decision maker (the person or organization for whom the study was instigated in the first place). All persons involved in the study must have a clear understanding of the study objectives and the limitations associated with the chosen physical sampling and analytical (measurement) techniques before anyone can make any sense of the resulting data.

The DQO Process is the Scientific Method

The DQO process is really just a formalization of sampling design and the scientific method. We will explain the DQO process in detail, but first we need to understand the concepts of population, sample, random sampling, and hypothesis test.

WHAT IS A POPULATION AND WHAT IS A SAMPLE?

In everyday language, the word “population” refers to all the people or organisms contained within a specific country, area, region, etc. When we talk about the population of the United States, we usually mean something like “the total number of people who currently reside in the U.S.”

In the field of statistics, however, the term *population* is defined operationally by the question we ask: it is the entire collection of measurements about which we want to make a statement (Zar, 1999, p. 16; Berthouex and Brown, 1994, p. 7; Gilbert, 1987, Chapter 2).

For example, if the question is “What does the concentration of dissolved oxygen look like in this stream?”, the question must be further refined until a suitable population can be defined: “What is the average concentration of dissolved oxygen in a particular section of a stream at a depth of 0.5 m over a particular 3-day period?” In this case, the population is the set of all possible measurements of dissolved oxygen in that section of the stream at 0.5 m within that time period. The section of the stream, the time period, the method of taking water samples, and the method of measuring dissolved oxygen all define the population.

A *sample* is defined as some subset of a population (Zar, 1999, p. 17; Berthouex and Brown, 1994, p. 7; Gilbert, 1987, Chapter 2). If the sample contains all the elements of the population, it is called a *census*. Usually, a population is too large to take a census, so a portion of the population is sampled. The statistical definition of the word sample (a selection of individual population members) should not be confused with the more common meaning of a *physical sample* of soil (e.g., 10g of soil), water (e.g., 5ml of water), air (e.g., 20 cc of air), etc.

RANDOM VS. JUDGMENT SAMPLING

Judgment sampling involves subjective selection of the population units by an individual or group of individuals (Gilbert, 1987, Chapter 3). For example, the number of samples and sampling locations might be determined based on expert opinion or historical information. Sometimes, public opinion might play a role and samples need to be collected from areas known to be highly polluted. The uncertainty inherent in the results of a judgment sample cannot be quantified and statistical methods cannot be applied to

judgment samples. Judgment sampling does *not* refer to using prior information and the knowledge of experts to define the area of concern, define the population, or plan the study. Gilbert (1987, p. 19) also describes “haphazard” sampling, which is a kind of judgment sampling with the attitude that “any sample will do” and can lead to “convenience” sampling, in which samples are taken in convenient places at convenient times.

Probability sampling or **random sampling** involves using a random mechanism to select samples from the population (Gilbert, 1987, Chapter 3). All statistical methods used to quantify uncertainty assume some form of random sampling has been used to obtain a sample. At the simplest level, a simple random sample is used in which each member of the population has an equal chance of being chosen, and the selection of any member of the population does not influence the selection of any other member. Other probability sampling methods include stratified random sampling, composite sampling, and ranked set sampling. We will discuss these methods in detail later in this chapter.

THE HYPOTHESIS TESTING FRAMEWORK

Every decision you make is based on a set of assumptions. Usually, you try to make the “best” decision given what you know. The simplest decision involves only two possible choices, with the choice you make depending only on whether you believe one specific condition is true or false. For example, when you get up in the morning and leave your home, you have to decide whether to wear a jacket or not. You may make your choice based on whether you believe it will rain that day. Table 2.1 displays the framework for your decision-making process.

Your Decision	What Happens	
	It Does Not Rain	It Rains
Wear Jacket	Mistake I	Correct Decision
Leave Jacket	Correct Decision	Mistake II

Table 2.1 Hypothesis testing framework for wearing a jacket

If you wear your jacket and it rains, you made a “correct” decision, and if you leave your jacket at home and it does not rain then you also made a “correct” decision. On the other hand, if you wear your jacket and it does not rain, then you made a “mistake” in some sense because you did not need to wear your jacket. Also, if you leave your jacket at home and it does rain,

then you also made a “mistake.” Of course, you may view the “cost” of making mistake II as greater than the cost of making mistake I.

You can think of the above example as an illustration of a hypothesis test. We will say the null hypothesis is that it will not rain today. You then gather information from the television, Internet, radio, newspaper, or personal observation and make a decision based on this information. If the null hypothesis is true and you decide not to leave your jacket at home, then you made the “correct” decision. If, on the other hand, the null hypothesis is false and you decide to leave your jacket at home, you have made a mistake.

Decisions regarding the environment can also be put into the hypothesis testing framework. Table 2.2 illustrates this framework in the context of deciding whether contamination is present in the environment. In this case, the null hypothesis is that no contamination is present.

Your Decision	Reality	
	No Contamination	Contamination
Contamination	Mistake: Type I Error (Probability = α)	Correct Decision (Probability = $1-\beta$)
No Contamination	Correct Decision	Mistake: Type II Error (Probability = β)

Table 2.2 Hypothesis testing framework for deciding on the presence of contamination in the environment when the null hypothesis is “no contamination”

Statisticians call the two kinds of mistakes you can make a **Type I error** and a **Type II error**. Of course, in the real world, once you make a decision, you take an action (e.g., clean up the site or do nothing), and you hope to find out eventually whether the decision you made was the correct decision.

For a specific decision rule, the probability of making a Type I error is usually denoted with the Greek letter α (alpha). This probability is also called the **false positive rate**. The probability of making a Type II error is usually denoted with the Greek letter β (beta). This probability is also called the **false negative rate**. The probability $1-\beta$ denotes the probability of correctly deciding there is contamination when in fact it is present. This probability is called the **power** of the decision rule. We will talk more about the hypothesis testing framework in Chapters 6 and 7, and we will talk about power in Chapter 8.

COMMON MISTAKES IN ENVIRONMENTAL STUDIES

The most common mistakes that occur in environmental studies include the following:

- **Lack of Samples from Proper Control Populations.** If one of the objectives of an environmental study is to determine the effects of a pollutant on some specified population, then the sampling design must include samples from a proper control population. This is a basic tenet of the scientific method. If control populations were not sampled, there is no way to know whether the observed effect was really due to the hypothesized cause, or whether it would have occurred anyway.
- **Using Judgment Sampling to Obtain Samples.** When judgment sampling is used to obtain samples, there is no way to quantify the precision and bias of any type of estimate computed from these samples.
- **Failing to Randomize over Potentially Influential Factors.** An enormous number of factors can influence the final measure associated with a single sampling unit, including the person doing the sampling, the device used to collect the sample, the weather and field conditions when the sample was collected, the method used to analyze the sample, the laboratory to which the sample was sent, etc. A good sampling design controls for as many potentially influencing factors as possible, and randomizes over the factors that cannot be controlled. For example, if there are four persons who collect data in the field, and two laboratories are used to analyze the results, you would not send all the samples collected by persons 1 and 2 to laboratory 1 and all the samples collected by persons 3 and 4 to laboratory 2, but rather send samples collected by each person to each of the laboratories.
- **Collecting Too Few Samples to Have a High Degree of Confidence in the Results.** The ultimate goal of an environmental study is to answer one or more basic questions. These questions should be stated in terms of hypotheses that can be tested using statistical procedures. In this case, you can determine the probability of rejecting the null hypothesis when in fact it is true (a Type I error), and the probability of not rejecting the null hypothesis when in fact it is false (a Type II error). Usually, the Type I error is set in advance, and the probability of correctly rejecting the null hypothesis when in fact it is false (the power) is calculated for various sample sizes. Too often, this step of determining power and sample size is neglected, resulting in a study from which no conclusions can be drawn with any great degree of confidence.

Following the DQO process will keep you from committing these common mistakes.

THE DATA QUALITY OBJECTIVES PROCESS

The Data Quality Objectives (DQO) process is a systematic planning tool based on the scientific method that has been developed by the U.S. Environmental Protection Agency (USEPA, 1994a). The DQO process provides an easy-to-follow, step-by-step approach to decision-making in the face of uncertainty. Each step focuses on a specific aspect of the decision-making process. Data Quality Objectives are the qualitative and quantitative statements that:

- Clarify the study objective.
- Define the most appropriate type of data to collect.
- Determine the most appropriate conditions under which to collect the data.
- Specify acceptable levels of decision errors that will be used as the basis for establishing the quantity and quality of data needed to support the decision.

Once the DQOs are specified, it is the responsibility of the team to determine the most cost-effective sampling design that meets the DQOs. A sampling design or sampling plan is a set of instructions to use to scientifically investigate the study objective and come up with a quantifiable answer. We will discuss several commonly used sampling designs later in this chapter, but it is important to note that the actual method of analysis/estimation is part of the design, and various available choices must be considered to choose the most resource-effective combination of the method of sampling and the method of analysis/estimation.

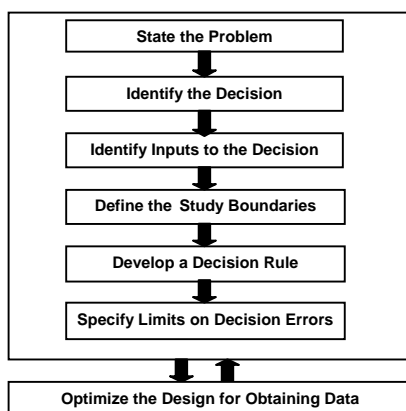


Figure 2.1 The Data Quality Objectives (DQO) Process

Figure 2.1 displays a flowchart of the DQO process. The DQO process is described in detail in USEPA (1994a). The DQO process is a series of seven planning steps, as outlined below.

1. **State the Problem.** Clearly and concisely define the problem so that the focus of the study will be unambiguous. Activities include:
 - Identify members of the planning team.
 - Identify the primary decision maker and each team member's role and responsibilities.
 - Develop a concise description of the problem by describing what is causing the problem and reviewing prior studies and existing information to gain a better understanding of the problem.
 - Specify available resources and relevant deadlines for the study.
2. **Identify the Decision.** Define the decision statement that the study will attempt to resolve. Activities include:
 - Identify the principal study question. For example: Is the concentration of 1,2,3,4-tetrachlorobenzene (TcCB) in the soil at a specific site "significantly" above "background" levels?
 - Define possible (alternative) actions to take based on the answer to the principal study question. For example, if the concentration of TcCB is significantly above background level, then require remediation; otherwise, do nothing.
 - Combine the principal study question and the alternative actions into a decision statement.
 - Organize multiple decisions.
3. **Identify Inputs to the Decision.** Identify the information that needs to be obtained and the measurements that need to be taken to resolve the decision statement. Activities include:
 - Identify the information required to resolve the decision statement. For example, what is the distribution of TcCB concentrations at the site of concern, and what is the distribution of TcCB concentrations for a "background" site?
 - Determine the sources for each item of information. Sources may include previous studies, scientific literature, expert opinion, and new data collections.
 - Identify the information that is needed to establish the action level. For example, will the "background" level of TcCB be established based on sampling a nearby "control" site or based on a regulatory standard?

- Confirm that appropriate analytical methods exist to provide the necessary data. Develop a list of potentially appropriate measurement methods, noting the method detection limit and limit of quantitation for each method.
4. **Define the Study Boundaries.** Define the spatial and temporal boundaries that are covered by the decision statement. A clear connection should be made between the study boundaries and the statement of the problem so that the decisions are relevant to the problem stated in Step 1. Steps 3 and 4 involve defining the population(s) of interest, how it (they) will be sampled, and how the physical samples will be measured. Activities for Step 4 include:
- Specify the characteristics that define the population of interest.
 - Define the geographic area within which all decisions apply.
 - When appropriate, divide the population into relatively homogeneous strata.
 - Determine the timeframe within which all decisions apply.
 - Determine when to collect the data.
 - Define the scale of decision making. For example, will only one decision be made for the whole site, or will the site be divided into smaller sub-areas and a separate decision made for each sub-area?
 - Identify any practical constraints on data collection.
5. **Develop a Decision Rule.** Define the statistical parameter of interest, specify the action level, and integrate the previous DQO outputs into a single statement that describes the logical basis for choosing among alternative actions. Activities include:
- Specify one or more statistical parameter(s) that characterize(s) the population and that are most relevant to the decision statement. For example, the average concentration of TcCB, the median concentration of TcCB, and the 95th percentile of the concentration of TcCB.
 - Specify the action level(s) for the study.
 - Combine the outputs from the previous DQO steps into one or more “If ...then...” decision rules that define the conditions that would cause the decision maker to choose among alternative actions. For example, “If the average concentration of TcCB in a sub-area is greater than the background average concentration then remediate the sub-area.”
6. **Specify Tolerable Limits on Decision Errors.** Define the decision maker’s tolerable decision error rates based on a consideration of the consequences of making an incorrect decision. Since decisions are

made based on a sample rather than a census, and there is uncertainty in the measurements, incorrect decisions can be made. The probabilities of making incorrect decisions are referred to as decision error rates. Activities for this step include:

- Determine the full possible range of the parameter of interest.
- Identify the decision errors and formulate the null hypothesis.
- Specify a range of parameter values where the consequences of decision errors are relatively minor (gray region).
- Assign probability values to points above and below the action level that reflect the tolerable probability for the occurrence of decision errors.

Example of a Decision Performance Goal Diagram

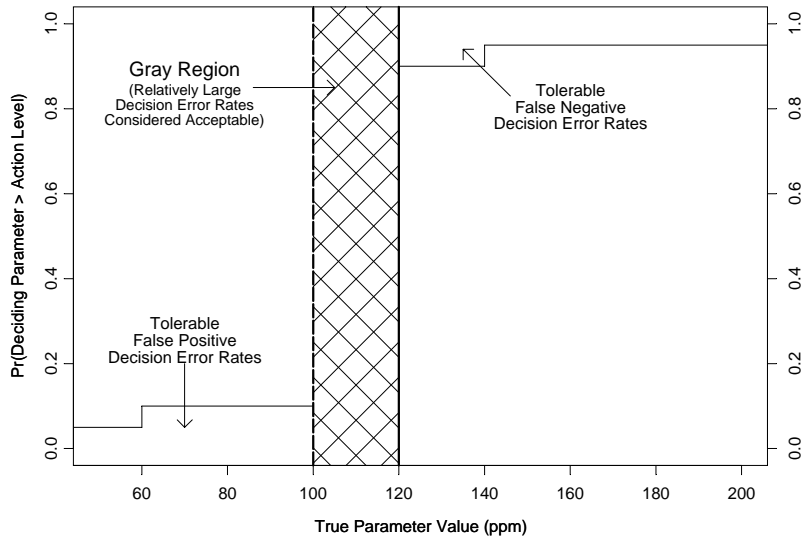


Figure 2.2 Decision performance goal diagram for the null hypothesis that the parameter is less than the action level of 100 ppm

Figure 2.2, similar to Figure 6-2 of USEPA (1994a, p. 36), illustrates a decision performance goal diagram in which the full range of the parameter is assumed to be between about 50 and 200 ppm. The null hypothesis is that the true value of the parameter is less than the action level of 100 ppm. The tolerable false positive decision error rate is 5% as long as the true value is less than 60 ppm. If the true value is between 60 and 100 ppm, then the tolerable false positive rate is 10%. If the true value is between 100 and 120 ppm, this is a gray region

where the consequences of a decision error are relatively minor. If the true value is between 120 and 140 ppm, the tolerable false negative decision error rate is 10%, and if the true value is bigger than 140 ppm the tolerable false negative decision error rate is 5%.

7. **Optimize the Design.** Evaluate information from the previous steps and generate alternative sampling designs. Choose the most resource-efficient design that meets all DQOs. Activities include:
 - Review the DQO outputs and existing environmental data.
 - Develop several possible sampling designs.
 - Formulate the mathematical expressions required to solve the design problems for each design alternative. These expressions include the relationship between sample size and decision error rates, and the relationship between sample size and cost of the design.
 - Select the optimal sample size that satisfies the DQOs for each possible sampling design.
 - Select the most resource-effective sampling design that satisfies all of the DQOs.
 - Document the operational details and theoretical assumptions of the selected sampling design in the Sampling and Analysis Plan.

Example of a Power Curve

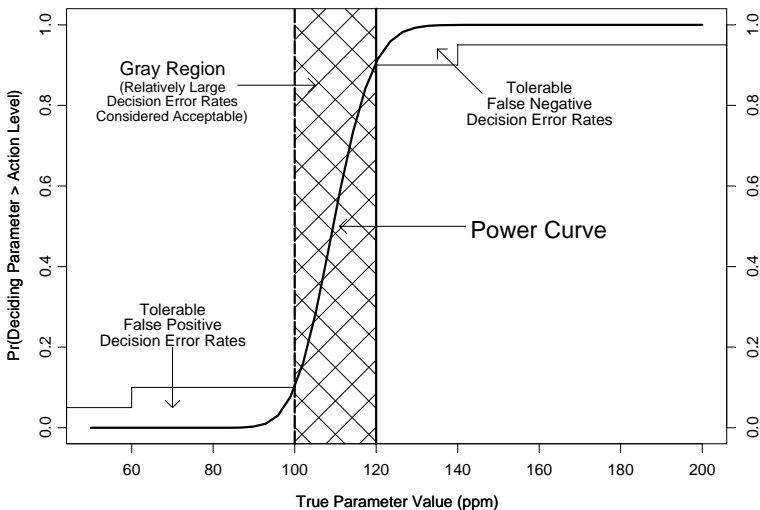


Figure 2.3 Example of a power curve for the null hypothesis that the parameter is less than the action level of 100 ppm

Figure 2.3, similar to Figure 7-1 of USEPA (1994a, p. 40) illustrates a power curve for a particular sampling design drawn on top of the decision performance goal diagram shown in Figure 2.2. Here you can see that this particular sampling design satisfies the performance goals.

Steps 3 and 4 should include developing a **quality assurance project plan (QAPP)** to document **quality assurance (QA)** and **quality control (QC)** and insure the integrity of the final results (USEPA, 1998b). A good QAPP covers instructions for sample collection in the field, handling, laboratory analysis and reporting, data coding, statistical analyses, and reports. Embedded in these instructions is the **chain of custody procedures** for documenting who has custody of the samples and the current conditions of the samples from the point of collection in the field to the analysis at the laboratory. Chain of custody procedures are used to ensure that samples are not lost, tampered with, or improperly stored or handled. See Keith (1991) and USEPA (1998b) for more information.

Step 7 requires information from previous studies and/or a pilot study to quantify the amount of variability that is typical in samples. Once this information is available, you can estimate the required sample size based on the statistical method of analysis that will be used (see Chapter 8). Even if information on sample variability is available from previous studies, however, it is almost always advisable to conduct a pilot study in order to “fine tune” the QA/QC sampling plan and the overall sampling design. Sample size requirements should take into account that a certain proportion of the samples will be unusable due to loss, mislabeling, mishandling, or some other factor that keeps the samples from meeting the specified QA/QC standards.

The DQO process is iterative. During the design optimization step (Step 7), you may discover that there are not enough funds and/or staff available to answer the question, given the required decision error rates specified in Step 6 and the required sample sizes determined in Step 7. In this case, you may need to adjust the budget of the study or the decision error rates, or investigate alternative methods of sampling that may yield smaller variability. Once a sampling design is chosen, you should develop a written protocol for implementing the sampling design and QAPP programs. See Gilbert (1987), Keith (1991), and USEPA (1989a; 1992b; 1994a,b; 1996) for more information on sampling design.

SOURCES OF VARIABILITY AND INDEPENDENCE

Figure 1.2 in Chapter 1 displays several sources of variability in measurements from environmental studies. Some potential sources of variability include:

- Natural variability over space.
- Natural variability over time (including seasonal and year-to-year fluctuations).
- Field sampling variability (sample collection, sample handling, and transportation).
- Within laboratory variability (day-to-day, machine-to-machine, technician-to-technician, etc.).
- Between laboratory variability.

When you are designing a sampling program, it is very important to be aware of these different sources of variability and to know how much they contribute to the overall variability of a measure. For example, RCRA regulations for groundwater monitoring at hazardous and solid waste sites require a minimum of only one upgradient well (see Figure 1.1 in Chapter 1). If there is substantial natural spatial variability in the concentration of some chemical of concern, then we may not be able to tell whether a difference in concentrations between the upgradient well and a downgradient well is due to actual contamination showing up at the downgradient well or simply due to natural spatial variability.

Independent vs. Dependent Observations

A key concept in statistical design and analysis is the idea of *independent observations*. Here is a non-technical definition: observations are independent of one another if knowing something about one observation does not help you predict the value of another observation. As an example of independent observations, suppose you have a standard deck of 52 playing cards, which you shuffle before picking a single card. You look at the card, put it back in the deck, shuffle the deck again, and pick another card. Knowing that the card you picked the first time was the 10 of diamonds will not help you predict what the next card will be when you pick from the deck again. The values of the two different cards that you pick are independent. On the other hand, if you had kept the 10 of diamonds after you picked it and not returned it to the deck, then you know that the next card you pick cannot be the 10 of diamonds. In this case, the two observations are dependent.

The idea of independence is closely linked to the idea of accounting for sources of variability. Suppose we are monitoring the concentration of arsenic in groundwater around a hazardous waste site and the flow of groundwater is extremely slow. If we sample say biweekly or monthly, then the temporal variability at a monitoring well will be small, so that observations taken close together in time will resemble one another more than observations taken further apart in time. Thus, if we combine monthly observations over 2 years, these 24 observations would not be considered to be independent of each other; they would exhibit temporal correlation (see Chapter 11).

We would have to adjust the sampling frequency to something like quarterly to try to produce observations that act like they are independent.

Similarly, if there is no appreciable natural spatial variability of arsenic concentrations over the area we are monitoring, then the combined observations from an upgradient well and a downgradient well could be considered to be independent (assuming there is no leakage from the waste site). On the other hand, if there is substantial spatial variability (but no leakage from the waste site), and the concentrations at the downgradient well tend to be larger than the concentrations at the upgradient well, then the combined observations from the upgradient well and the downgradient well would not be considered independent; knowing which well an observation comes from (upgradient or downgradient) helps us predict the value of the observation. If we know the average concentrations at the upgradient and downgradient wells (e.g., 5 ppb at the upgradient well and 10 ppb at the downgradient well), we can subtract these averages from the observations to produce “residual values.” The combined residual values would be considered independent, because knowing whether the residual value came from the upgradient or downgradient well does not help us predict the actual value of the residual value.

Most standard statistical methods assume the observations are independent. This is a reasonable assumption as long as we are careful about accounting for potential sources of variability and randomizing over potentially influential factors (e.g., making sure each laboratory analyzes samples collected by each different collector). Time series analysis, discussed in Chapter 11, and spatial statistics, discussed in Chapter 12, are special ways of accounting for variability over time and space. A very common method of accounting for physical sources of variability in the field is stratified random sampling, which is discussed in the next section.

METHODS OF RANDOM SAMPLING

This section describes four methods of random sampling: simple random sampling, systematic sampling, stratified random sampling, composite sampling, and ranked set sampling. Other methods of random sampling include two-stage and multi-stage random sampling, double random sampling, sequential random sampling, and adaptive random sampling. See Gilbert (1987), Keith (1991, 1996), Thompson (1992), and Cochran (1977) for more information.

In our discussions, we will explain the various methods of random sampling in English. To compare the methods of sampling, however, we need to use equations and talk about concepts like the population mean, sample mean, and variance of the sample mean. We will discuss these concepts in detail in Chapters 3 and 4. For now, you may want to skip the equations and return to them after you have read these later chapters.

Simple random sampling is, true to its name, the simplest type of sampling design. Recall that a sample is a subset of the population. A simple random sample (SRS) is obtained by choosing the subset in such a way that every individual in the population has an equal chance of being selected, and the selection of one particular individual has no effect on the probability of selecting another individual. The word *random* denotes the use of a probabilistic mechanism to ensure that all units have an equal probability of being selected, rather than the colloquial meaning “haphazard.”

In order to realize an SRS from a population, a complete listing of the units in the population may be needed. These units are referred to as **sampling units**. For example, suppose we are interested in determining the concentration of a chemical in the soil at a Superfund site. One way to define the population is as the set of all concentrations from all possible physical samples of the soil (down to some specified depth). In the DQO process, we may further refine our definition of the population to mean the set of all concentrations from all possible physical samples taken on a grid that overlays the site. The grid points are then the sampling units, and an SRS will consist of a subset of these grid points in which each grid point has an equal probability of being included in the sample. Figure 2.4 illustrates the idea of simple random sampling for a grid with $N = 160$ points for which $n = 16$ points were selected at random.

Example of SRS with $N=160$ and $n=16$

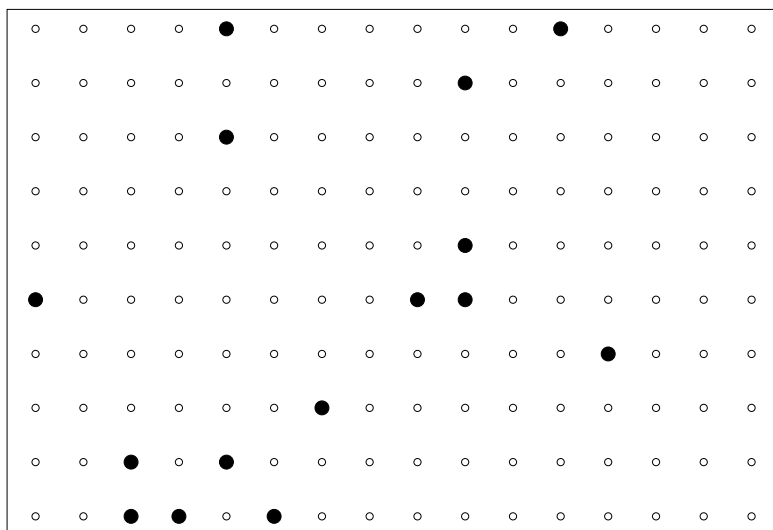


Figure 2.4 Illustration of simple random sampling on a grid

In Chapters 3 and 4 we will discuss the concepts of the population mean, population variance, sample mean, and sample variance. One way to compare sampling methods is based on the variance of the sample mean (how much wiggle it has). For simple random sampling, the population mean or average, denoted by μ , is estimated by the sample mean:

$$\hat{\mu}_{SRS} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.1)$$

where n denotes the number of sampling units selected in the SRS and x_i denotes the value of the measurement on the i^{th} sampling unit that was selected. The variance of the sample mean is given by:

$$\text{Var}(\hat{\mu}_{SRS}) = \text{Var}(\bar{x}) = \sigma_x^2 = \frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right) \quad (2.2)$$

where σ^2 denotes the population variance and N denotes the number of sampling units in the population (Cochran, 1977, p. 23; Gilbert, 1987, p. 28). Note that as the sample size n increases, the variance of the sample mean decreases. For a finite population, if we take a census so that $n = N$, the sample mean is the same as the population mean and the variance of the sample mean is 0. The quantity n/N in Equation (2.2) is called the **finite population correction factor**. For an infinite population where $N = \infty$, the finite population correction factor is 0 so that the variance of the sample mean depends only on the sample size n .

$$\text{Var}(\hat{\mu}_{SRS}) = \text{Var}(\bar{x}) = \frac{\sigma^2}{n} \quad (2.3)$$

Often in environmental studies, the finite population correction factor is set to 0 because the size of the population N is much, much larger than the sample size n .

Determining Grid Size

A common question is: “If I decide to overlay the area with a grid and define my population by the set of all possible points on the grid, how fine should the grid be? Should the points be spaced by 10 feet, one foot, half of a foot, or some other distance?” For simple random sampling, the answer is:

Make the grid of possible points to sample from as fine as possible, since you would like to extrapolate your results to the whole area. When you overlay the area with a grid, you are in effect taking a systematic sample of all possible sampling points (see the next section). You therefore may have introduced bias, because even if you sample from all of the points on the grid (instead of taking a random sample of points), the mean based on sampling all of the grid points may not equal the true population mean. As you make the grid finer, however, the mean of all grid points will get close to the true population mean.

Systematic Sampling

Systematic sampling involves choosing a random starting point, and then sampling in a systematic way, for example, along a line every 2 feet, or on a square or triangular grid. Figure 2.5 illustrates systematic sampling on a triangular grid. In this figure, the coordinates of the sampling point in the lower left-hand corner were generated by random numbers; once this coordinate was determined, the other sampling coordinates were completely specified.

Example of Systematic Sample on a Triangular Grid

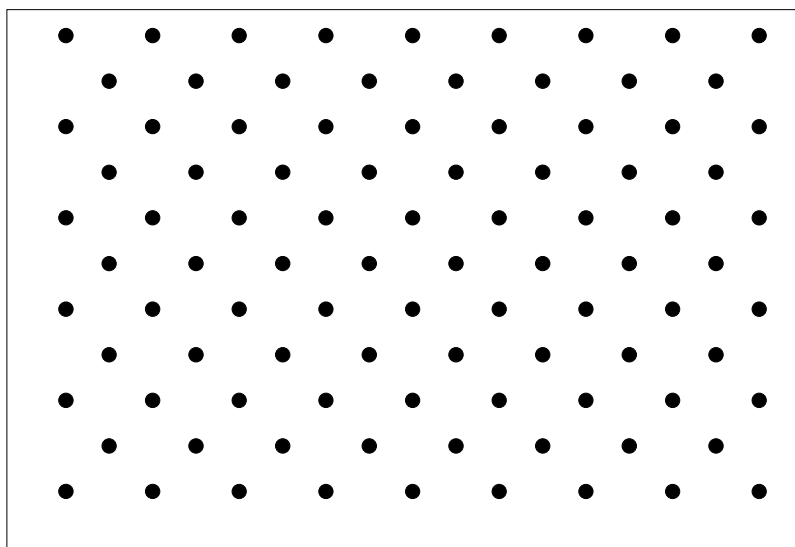


Figure 2.5 Illustration of systematic sampling with a triangular grid

Systematic sampling is useful when you are trying to uncover “hot spots” of highly contaminated areas. Gilbert (1987, Chapter 10) discusses how to determine grid size when using square, rectangular, or triangular grids to

search for hot spots. Elipgrid-PC is software for determining grid spacing to detect hotspots with a specified probability. It was originally developed at Oak Ridge National Laboratory and is available at the following URL: <http://etd.pnl.gov:2080/DQO/software/elipgrid.html>. Visual Sample Plan (VSP; see <http://terrassa.pnl.gov:2080/DQO/software/vsp>) is an updated version of Elipgrid-PC that also includes sample size calculations for simple random sampling.

If you are trying to estimate a mean or a total, systematic sampling works as well as or better than simple random sampling if there are no trends or natural strata in the population you are sampling. In general, however, it is difficult to obtain reliable estimates of variability with systematic sampling, and if some kind of natural trend or pattern is present, systematic sampling may yield biased estimates (Gilbert, 1987, Chapter 10).

Stratified Random Sampling

Another type of sampling design is called stratified random sampling. In stratified random sampling, the population (site or process) is divided into two or more non-overlapping strata, sampling units are defined within each stratum, then separate simple random or systematic samples are chosen within each stratum. Population members within a stratum are thought to be in some way more similar to one another than to population members in a different stratum. If this is in fact the case, then the sample mean based on stratified random sampling is less variable than the sample mean based on simple random sampling. Figure 2.6 illustrates the idea of stratified random sampling.

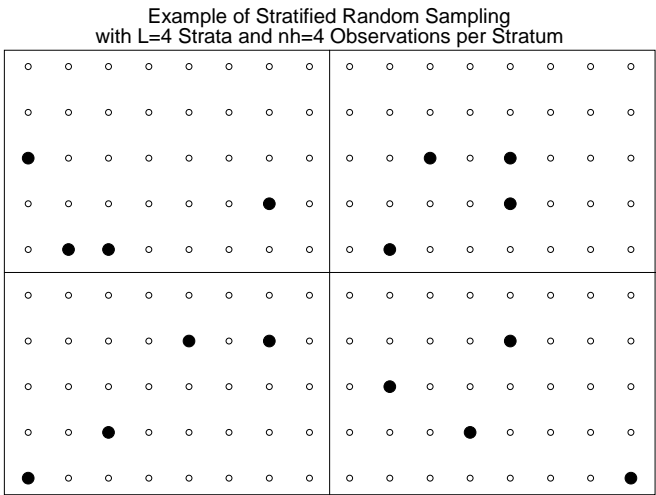


Figure 2.6 Illustration of stratified random sampling

It is important to note that the formulas for estimators such as the sample mean or proportion and the associated confidence intervals (see Chapter 5) need to be modified to be applicable to data from stratified random sampling. For stratified random sampling, the population mean of the h^{th} stratum is estimated by choosing n_h sampling units and computing the sample mean:

$$\hat{\mu}_h = \bar{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi} \quad (2.4)$$

where x_{hi} denotes the value of the measurement on the i^{th} sampling unit that was selected in stratum h . The population mean over all strata is estimated by a weighted average of the sample means from each stratum:

$$\hat{\mu}_{Stratified} = \sum_{h=1}^L W_h \hat{\mu}_h = \sum_{h=1}^L W_h \bar{x}_h \quad (2.5)$$

where

$$W_h = \frac{N_h}{N} \quad (2.6)$$

$$N = \sum_{h=1}^L N_h \quad (2.7)$$

L denotes the total number of strata, N_h denotes the number of sampling units in the h^{th} stratum, and N denotes the total number of sampling units (Cochran, 1977, p. 91; Gilbert, 1987, p. 46). If we let n denote the total sample size, then

$$n = \sum_{h=1}^L n_h \quad (2.8)$$

The estimator of the population mean given in Equation (2.5) is *not* the same as the simple sample mean taken over all observations in all the strata unless $n_h/N_h = n/N$ for all of the L strata, that is, the sampling fraction is the same in all of the strata, which is called **proportional allocation**.

The variance of the stratified sample mean is given by:

$$\begin{aligned} \text{Var}(\hat{\mu}_{\text{Stratified}}) &= \sum_{h=1}^L W_h^2 \text{Var}(\bar{x}_h) \\ &= \sum_{h=1}^L W_h^2 \frac{\sigma_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \end{aligned} \quad (2.9)$$

where σ_h^2 denotes the variance of the h^{th} stratum (Cochran, 1977, p. 92; Gilbert, 1987, p. 47). Cochran (1977, pp. 96–99) and Gilbert (1987, pp. 50–52) show how to minimize this variance if you know the variability within each stratum and the cost of taking a sample within each stratum.

When Should You Use Stratified Random Sampling?

The main advantage of the stratified random sampling design is its ability to provide a greater coverage of the population. By choosing sampling units from all strata, you are avoiding the possibility that all of the sampled units may come from the same or adjacent geographical areas and certain parts of the population are not represented. Stratified sampling is beneficial if the population you are sampling is fairly heterogeneous and you have a good idea of how to divide the population up into strata that are fairly homogeneous.

Composite Sampling

So far in our discussions of simple and stratified random sampling, we have inherently assumed that the physical samples are measured only once (e.g., a sample of groundwater is pumped from a monitoring well and the concentration of arsenic in that sample is measured). Sometimes, you may want to take two samples very close together in space or time, sometimes you may want to split a physical sample into two or more subsamples and take a measure on each, or sometimes you may want the laboratory to perform two or more analyses on a single physical sample. Each of these is an example of a **replicate**, so you must be very explicit about what you mean when you use this term. Replicates are used in QA/QC studies to estimate within-sample variability. If the within-sample variability is fairly large

(e.g., of the same order as between-sample variability), and the cost of analysis is fairly small compared to the cost of sample collection, you may want to specify two or more replicates in the sampling design.

Often, however, for environmental studies the cost of collecting a sample is relatively small compared to the cost of analyzing it. For instance, in sampling from a site which has highly radioactive material, the entire crew needs to rehearse the sample collection procedure to minimize the exposure and the time spent at the site for sampling. Once the sampling crew is in the field taking soil samples, however, the additional cost of collecting more soil samples is relatively small compared to the cost of analyzing the soil samples for radioactivity. In such situations, cost-effective sampling designs can be achieved by composite sampling (Lovison et al., 1994; USEPA, 1995a).

Composite samples are obtained by physically mixing the material obtained from two or more sampling units. Then measurements are obtained by analyzing subsamples (replicates) from each composite sample. Figure 2.7 illustrates composite sampling in which a composite sample is created by mixing together three grab samples (sampling units). A total of four composite samples are created from 12 grabs. Two subsamples (replicates) are taken from each of the four composite samples and measured, yielding a total of eight measurements.

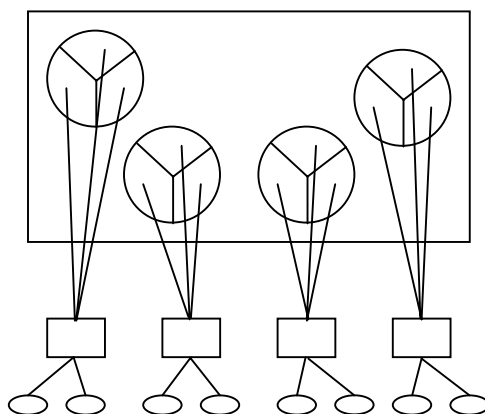


Figure 2.7 Illustration of composite sampling in which three grabs are mixed together to form a composite sample, then two subsamples are taken from the composite sample

Compositing simply represents a physical rather than mathematical mechanism of averaging the measurements from individual sampling units that make up the composite. When the cost of collecting samples is small, a large number of samples may be collected to ensure a large coverage of the population.

Compositing is a cost-effective means of estimating population means. Estimating the uncertainty and obtaining confidence intervals are not always possible with measurements from composite samples since information regarding the extremes is lost when grab samples are composited. Compositing can also be used to efficiently identify “hot-spots” if the sampled material does not deteriorate in storage. In this case, the individual units are strategically composited in groups and the “hot” samples are identified based on measurements from the composite (Sengupta and Neerchal, 1994). It is also possible to apply compositing with stratification or other types of designs as a screening mechanism.

Composite Sampling vs. Simple Random Sampling

Suppose we take n grab samples and divide them evenly into g groups of size h (so $h = n/g$), then composite the grabs within each group to produce g composite samples. We then take r subsamples (replicates) from each composite sample for a total of gr measurements. In Figure 2.7, we have $n = 12$, $g = 4$, $h = 3$, and $r = 2$.

Let x_{ij} denote the measurement from the j^{th} grab sample in the i^{th} group ($i = 1, 2, \dots, g$; $j = 1, 2, \dots, h$). For simple random sampling, our estimate of the population mean is the sample mean based on the n grab samples:

$$\bar{x} = \frac{1}{gh} \sum_{i=1}^g \sum_{j=1}^h x_{ij} \quad (2.10)$$

This is an unbiased estimator of the population mean μ , and the variance of this estimator is given by:

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{gh} \quad (2.11)$$

where σ^2 denotes the population variance of the grab samples:

$$\sigma^2 = \text{Var}(x_{ij}) \quad (2.12)$$

Let y_{ik} denote the k^{th} subsample (replicate) taken from the i^{th} composite sample (group), where $i = 1, 2, \dots, g$ and $k = 1, 2, \dots, r$. The average of all of the subsamples is:

$$\bar{y} = \frac{1}{gr} \sum_{i=1}^g \sum_{k=1}^r y_{ik} \quad (2.13)$$

The i^{th} composite sample is formed by physically mixing the h grab samples within the i^{th} group, so this is a mechanical way to produce a weighted average of the h grab samples within that group. We can therefore write y_{ik} as:

$$y_{ik} = \sum_{j=1}^h w_{ijk} x_{ij} \quad (2.14)$$

where $0 \leq w_{ij} \leq 1$ and

$$\sum_{j=1}^h w_{ijk} = 1 \quad (2.15)$$

The weights w_{ijk} represent the contribution of the j^{th} grab sample to the i^{th} composite. These weights are assumed to be random with an expected value of $1/h$ and a variance of σ_w^2 . In the case of perfect mixing for the i^{th} composite, each weight is exactly equal to $1/h$, the variance of the weights is 0 ($\sigma_w^2 = 0$), and the values of all of the r subsamples are exactly the same as the average of the h grab samples comprising the i^{th} composite group (assuming no measurement error). It is important to note that subsample measurements coming from the same composite ($y_{i1}, y_{i2}, \dots, y_{ir}$) are correlated because they are based on the same set of h grab samples. However, measurements from two different composites are uncorrelated because different composites are based on distinct sets of grab samples. It can be shown that the average of all the subsamples given in Equation (2.13) is an unbiased estimator of the population mean with variance given by:

$$\text{Var}(\bar{y}) = \frac{\sigma^2}{gh} + \frac{h}{gr} \sigma^2 \sigma_w^2 \quad (2.16)$$

How do you know when it is a good idea to use composite sampling? One way to decide is by comparing the variance of the estimator of the popu-

lation mean under simple random sampling with the variance of the estimator under composite sampling. Comparing Equations (2.11) and (2.16), we see that the variance of the estimator based on compositing can never be smaller than the variance of the estimator based on measuring each of the individual grab samples.

Let us consider the case now where you have n grab samples but your budget only allows you to measure r of the n grab samples, where $r < n$. Is it better to measure r of the n grab samples and compute the mean based on these observations, or is it better to combine all n grab samples into one composite, take r subsamples, and compute the mean based on these observations? In the first case, the variance of the sample mean is given by:

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{r} \quad (2.17)$$

In the second case, we have $g = 1$ and $h = n$, so based on Equation (2.16) the variance of the mean of all of the subsamples is given by:

$$\text{Var}(\bar{y}) = \frac{\sigma^2}{n} + \frac{n}{r} \sigma^2 \sigma_w^2 \quad (2.18)$$

The ratio of these two variances is given by:

$$\frac{\text{Var}(\bar{y})}{\text{Var}(\bar{x})} = \frac{r}{n} + n\sigma_w^2 \quad (2.19)$$

Composite sampling will be better than measuring r of the n grab samples when this ratio is less than 1, which will happen if σ_w^2 is sufficiently small (i.e., the grabs are well mixed in the composite).

Designing a sampling plan that may include compositing samples requires you to determine the number of grab samples (n), the number of composite groups (g) or the number of grab samples per group (h), and the number of subsamples per composite (r). Thorough mixing reduces σ_w^2 and makes composite sampling more competitive with simple random sampling.

Using Composite Sampling to Detect “Hot Spots”

As mentioned earlier, information regarding the extremes is lost when grab samples are composited. If, however, the grabs can be stored without any physical deterioration and are available for measurement at a later time, then we can use compositing to efficiently determine which, if any, grab samples exceed a given threshold without necessarily having to measure each individual grab sample. We shall give a simple example and refer the reader to Sengupta and Neerchal (1994) for more details.

Suppose we have $n = 7$ grab samples, identified by 1,2,...7, and we want to identify any grab sample that exceeds some concentration level L , a level that is deemed to be of concern (e.g., a soil screening level as described in Chapter 1). We can use the following procedure.

- 1. Make one composite consisting of all seven grabs. If this composite measurement does not exceed $L/7$, then declare all grabs to be “safe.” If the measurement exceeds $L/7$, conclude that at least one of the grabs is contaminated and go to Step 2. For this example, we will assume that if the composite measure is greater than $L/7$, it is also less than $2(L/7)$, so that we can conclude that at most one of the grabs is “contaminated.”
- 2. Form 4 composites as follows:

Composite	Constituent Grabs
A	1, 3, 5, 7
B	2, 3, 6, 7
C	4, 5, 6, 7

Measure composites A, B, and C and compare each measurement to $L/4$.

- 3. Note that each of A, B, and C could either be above $L/4$ (denoted by a +) or below $L/4$ (denoted by a -). The table below lists the eight possible outcomes and which grab this outcome indicates as being contaminated.

A	B	C	Contaminated Grab
+	-	-	1
-	+	-	2
+	+	-	3
-	-	+	4
+	-	+	5
-	+	+	6
+	+	+	7
-	-	-	None

As an example, suppose that the grab measurements are 1, 3, 7, 29, 4, 9 and 2 ppm. For this method, we do not measure each grab sample but instead make one composite of all the seven grabs. Assuming perfect mixing ($\sigma_w = 0$), the composite measurement is simply the average of the grab measurements, which in this case is 7.86 ppm. Now suppose that a sample is considered “hot” if it exceeds $L = 28$ ppm, so we will compare the concentration of the composite sample to $L/7 = 28/7 = 4$ ppm. Since the composite measurement is between 4 ppm and 8 ppm, we conclude that there is at most one “hot” grab. Thus we form three more composite as in Step 2. The resulting measurements are shown in Table 2.3.

Composite	Constituent Grabs	Measurement (ppm)
A	1,3,5,7	3.5
B	2,3,6,7	5.25
C	4,5,6,7	11

Table 2.3 Example of using compositing to determine which grab samples are “hot”

Comparing these measurements to 7 ppm, we look up (-,-,+) in the table shown in Step 3 above and conclude that grab 4 is possibly “hot.” One more measurement on grab 4 will reveal that it is in fact contaminated. Note that we made a total of five measurements instead of seven. For the procedure illustrated here we assume perfect mixing and no measurement errors. Sen-gupta and Neerchal (1994) show that for a number of situations this procedure leads to savings over testing every grab sample.

Ranked Set Sampling

Ranked Set Sampling (RSS) was first introduced by McIntyre (1952) and is described in Patil et al. (1994b), Johnson et al. (1996), and USEPA (1995b). In RSS, a large number of sampling units are selected from the population and only a subset of those are actually measured. Ranking based on an auxiliary variable is used to determine which of the selected sampling units are measured, hence its name. RSS is worth considering if

- The cost of measurements (lab analyses) is far greater than the cost of collecting the samples.
- An auxiliary characteristic is available that is highly correlated with the main characteristic of interest and is also inexpensive to measure.

We will first describe the procedure to obtain a ranked set sample and then discuss its advantages and disadvantages.

Steps to Create a Ranked Set Sample

In this section we will illustrate how to create a ranked set sample of size $m = 3$ objects, and then explain how to extend this to general values of m .

1. First collect three simple random samples, each of size 3. Let S_{ij} denote the j^{th} sampling unit within the i^{th} SRS ($i = 1, 2, 3$; $j = 1, 2, 3$).

SRS	Sampling Units		
1	S_{11}	S_{12}	S_{13}
2	S_{21}	S_{22}	S_{23}
3	S_{31}	S_{32}	S_{33}

2. Next, order the sampling units within each SRS in increasing order according to the values for the auxiliary variable. Let $S_{i(j)}$ denote the j^{th} ordered (ranked) sampling unit within the i^{th} SRS ($i = 1, 2, 3$; $j = 1, 2, 3$).

SRS	Sampling Units		
1	$S_{1(1)}$	$S_{1(2)}$	$S_{1(3)}$
2	$S_{2(1)}$	$S_{2(2)}$	$S_{2(3)}$
3	$S_{3(1)}$	$S_{3(2)}$	$S_{3(3)}$

3. Create the ranked set sample by selecting the j^{th} ranked sampling unit for the j^{th} SRS ($j = 1, 2, 3$). That is, the ranked set sample is $S_{1(1)}$, $S_{2(2)}$, and $S_{3(3)}$ (the diagonal in the table from the upper left to the lower right). These sampling units are shown in boldface in the table above. The actual variable of interest is measured only on these units, and we will denote these measurements by $x_{1(1)}$, $x_{2(2)}$, and $x_{3(3)}$.

In general, a RSS of size m requires taking m simple random samples each of size m (a total of m^2 sampling units).

Ranking a large number of sampling units is difficult and prone to error, especially if the ranking has to be done visually. It is therefore recommended that the size of each SRS that needs to be ranked should be no larger than three to five sampling units. You can increase the sample size of your ranked set sample by replicating the process r times. For example, if you want a total of $n = 15$ sampling units in your ranked set sample, simply create $r = 5$ ranked set samples, each of size $m = 3$. This will require a total of $5 \times 3^2 = 45$ sampling units, as opposed to $15^2 = 225$ sampling units.

The Auxiliary Variable

As we discussed earlier in the context of composite sampling, often the laboratory analyses are much more expensive than the cost of collecting samples (per unit). The auxiliary characteristic used for ranking, on the other hand, should be inexpensive to measure. Often, the ranking can be done visually. For example, if we are interested in estimating the average biomass volume in a forest, a ranking of small, medium, and large can be done visually, even though measuring the biomass volume will involve time-intensive analyses. In a contaminated site, visible soil characteristics may provide a means of ranking the units while a soil sample will have to be sent to the lab to actually obtain measurements.

Ranked Set Sampling vs. Simple Random Sampling: Estimating the Mean

The sample mean based on RSS is an unbiased estimator of the population mean μ . If we create r ranked set samples, each of size m , to produce a ranked set sample of size $n = rm$, then the variance of the sample mean based on RSS is

$$Var(\bar{x}_{RSS}) = \frac{\sigma^2}{n} \left[1 - \frac{1}{m\sigma^2} \sum_{i=1}^m (\mu_{(i)} - \mu)^2 \right] \quad (2.20)$$

where $\mu_{(i)}$ denotes the expected value of the i^{th} order statistic from a random sample of size m (Patil et al., 1994b). Comparing Equation (2.20) with Equation (2.3), we see that the sample mean based on RSS has a smaller variance than the sample mean based on a SRS of the same size. Thus, fewer samples are needed to achieve the same precision, leading to a savings in sampling costs. It is worth emphasizing, however, that to obtain unbiased estimates of variances you need two or more cycles (see Stokes, 1980 and Bose and Neerchal, 1998). Mode et al. (1999) discuss when ranked set sampling is cost effective based on considering the cost of measuring the auxiliary variable, the cost of ranking, and the cost of measuring the variable of interest.

Equation (2.20) assumes there is perfect correlation between the auxiliary variable and the variable of interest, and that there are no errors in ranking based on the auxiliary variable. Nevertheless, the variance of the sample mean based on RSS is always less than or equal to the variance of the sample mean based on SRS. If there is no correlation between the auxiliary variable and the variable of interest, then RSS will simply produce the same results as SRS.

Ranked Set Samples Are More Regularly Spaced Than Simple Random Samples

We conclude our discussion of RSS by illustrating a very important property of the RSS design with a simple example. Suppose we want to estimate the average weight of a herd of elephants. Furthermore, assume this simple herd of elephants has one calf for every mother elephant and has no father elephants at all. We will consider two options:

- **Simple Random Sampling.** Pick two elephants at random, weigh them, and use their average weight as an estimate of the average weight for the entire herd. Note that this procedure gives an unbiased estimate.
- **Ranked Set Sampling.** Pick two elephants randomly in the morning, pick the smaller of the two, and weigh it. Pick two elephants randomly in the afternoon, pick the larger of the two, and weigh it. Use the average weight of the two elephants chosen in this way as an estimate of the average weight for the entire herd. It can be shown that this estimate is also an unbiased estimate of the average weight of the herd.

Table 2.4 shows the results of SRS with $n = 2$ elephants, where c denotes a calf and M denotes a mother. For RSS, Table 2.5 shows the possible results for the morning sample (smaller of the two is chosen), Table 2.6 shows the possible results for the afternoon sample (larger of the two is chosen), and Table 2.7 shows the final results.

1 st Elephant	2 nd Elephant	
	c	M
c	cc	cM
M	cM	MM

Table 2.4 Possible results of SRS from the elephant herd with sample size $n = 2$

1 st Elephant	2 nd Elephant	
	c	M
c	c	c
M	c	M

Table 2.5 Possible results of the morning SRS in which two elephants are selected and then the smaller elephant is weighed

1 st Elephant	2 nd Elephant	
	<i>C</i>	<i>M</i>
	<i>C</i>	<i>M</i>
<i>M</i>	<i>M</i>	<i>M</i>

Table 2.6 Possible results of the afternoon SRS in which two elephants are selected and then the larger elephant is weighed

Morning Elephant	Afternoon Elephant			
	<i>C</i>	<i>M</i>	<i>M</i>	<i>M</i>
	<i>C</i>	<i>CC</i>	<i>CM</i>	<i>CM</i>
	<i>C</i>	<i>CC</i>	<i>CM</i>	<i>CM</i>
	<i>C</i>	<i>CC</i>	<i>CM</i>	<i>CM</i>
	<i>M</i>	<i>CM</i>	<i>MM</i>	<i>MM</i>

Table 2.7 Possible results of RSS for the elephant herd with final sample size $n = 2$

For SRS, the probability of getting two mothers in the sample (resulting in an overestimate of the mean) or two calves in the sample (resulting in an underestimate of the mean) is $2/4 = 50\%$. For RSS, the probability of getting two mothers in the sample or two calves in the sample is only $6/16 = 37.5\%$. Clearly RSS is superior to SRS in the sense that RSS gives a “representative” sample more often. It can be shown theoretically that RSS gives the more representative samples more often in general for any herd of elephants regardless of the proportion of calves in the herd (Lacayo and Neerchal, 1996).

The above example captures the salient feature of RSS of producing less variable and more “representative” samples. This is the reason why the sample mean from a RSS has a smaller variance than the sample mean from an SRS of the same size. See Patil et al. (1994b) and the references therein for a number of theoretical results comparing RSS to SRS for various standard estimation problems.

CASE STUDY

We end this chapter with a case study that illustrates using the DQO process to systematically develop a sampling plan to achieve a specific ob-

jective. For our case study, we are interested in determining the percentage of employees of a facility who have knowledge of specific critical facts related to emergency preparedness. Although the case study is based on a real project in which one of the authors was involved, all references to the actual facility and people have been removed. The case study is presented in terms of the seven steps of the DQO process. Neptune et al. (1990) present a case study of using the DQO process to design a remedial investigation/feasibility study (RI/FS) at a Superfund site.

1. State the Problem

A large facility deals with hazardous material on an everyday basis. The facility is required by a Federal law to impart Emergency Preparedness Training (EPT) to its employees every year. The facility has about 5,000 employees housed in approximately 20 buildings. In the past, the facility has provided training by requiring all employees to attend a training session. Assuming a conservative estimate of about 45 minutes of employee time per training session (including travel time and not counting the time of the training staff), the total time spent on EPT adds up to approximately $5,000 \times 0.75 = 3,750$ employee hours. At an average billing rate of \$80 per hour per employee, the cost of EPT coming out of overhead and research programs is about $3,750 \times \$80 = \$300,000$ per year. In the past, there has been no follow-up to verify that the employees have retained the material presented at the training session.

To save money and still comply with the law, the facility wants to change the training procedure as follows. Instead of each employee attending the formal training session, each employee will receive a list of emergency preparedness (EP) facts the general staff needs to know to be prepared for an emergency. A description of these EP facts is given below, listed in their order of importance (highest to lowest).

1. Emergency telephone number.
2. Location of the nearest fire alarm pull box.
3. Location of the Building Staging Area.
4. Building alarms and the corresponding required actions by staff members.

The facility wants to verify that the EP facts are retained by the employees during the year so that they can recall them should an emergency arise. The facility will ensure that a high percentage of employees in each facility know all of the above facts by actually testing a certain number of occupants of each building. Complete EPT will be required for all occupants of a building for which a prescribed proportion of the randomly selected employees fail to demonstrate the knowledge of these facts.

2. Identify the Decision

It is important that staff members who spend off-hours in the facility on a regular basis know each EP fact listed above without exception. However, it may be sufficient if only a high percentage of the remaining staff members know these facts. For example, it may be sufficient if 70% or more of the occupants of each building are familiar with the EP facts. We need to verify that a certain large proportion of the occupants of each building are familiar with the EP facts 1 through 4. For each building, if we decide that the proportion of occupants who know the EP facts is too small, then we will send everyone in that building to formal EP training.

3. Identify Inputs to the Decision

A complete listing of the staff and the respective locations within the facility is available. Training manuals containing the EP facts can be provided to each staff member and they can be interviewed either in person or by telephone to verify that they have knowledge of all four of the EP facts.

4. Define the Study Boundaries

Since the retention of EP information may not last very long, it is important to realize that this investigation needs to be repeated every year. A face-to-face interview or some honor code may need to be imposed to ensure that each staff member is in fact providing the verification based on memory rather than using notes.

5. Develop a Decision Rule

The number of employees who know the EP facts will be verified for each building. For each building, if 70% or more of the staff members who work in that building know the EP facts, we will declare the whole building is prepared for an emergency. If less than 70% of the staff members in a building know the EP facts, then everyone in the building will be asked to go through the formal EP training session.

6. Specify Acceptable Limits for the Decision Errors

Interviewing every single staff member to determine the true percentage who know EP facts 1 to 4 would be almost as time consuming as simply sending all of the employees to the EPT itself. Therefore, cost considerations dictate taking a representative sample of staff in each building and interviewing each person in the sample to estimate the percentage of staff in each building who know the EP facts. For each building, a random sample of its occupants will be chosen (the recommended sample sizes are given in the tables below). Each selected staff member will be interviewed to deter-

mine whether he/she knows facts 1 to 4 of the EP checklist. If the number of DUDs (Doesn't Understand Directions) in the sample does not exceed t (t is given in tables below), then the building can be declared to be emergency prepared. Otherwise, the building is considered unprepared for an emergency and all staff in the building will be required to go through EPT.

We assume that the verification is error-free for all the staff members included in the sample. The decision, however, is based on information from only a subset of the occupants of the building. Therefore, we must recognize that there are risks involved with our decision. Table 2.8 summarizes the two kinds of decision errors that can be made and the consequences of each.

Decision	True Percentage of DUDs	
	≤ 30%	> 30%
Train Occupants	Decision Error I; Waste of Money	Correct Decision
Building is Emergency Prepared	Correct Decision	Decision Error II; Compliance Issue

Table 2.8 Decision errors and their consequences for the EPT case study

Risks involved with the two kinds error shown in the table above are competing with each other. That is, when the probability of making Decision Error I goes down, the probability of making Decision Error II goes up, and vice-versa. The approach taken in developing a statistical decision rule is to hold the probability of making the most critical error to a predetermined low level and to do the best we can with lowering the probability of making the other error. In this example, it is clear that controlling the probability of making the error leading to compliance issues (Decision Error II) should be a higher priority than controlling the probability of making the error leading to a waste of money (Decision Error I).

In a series of meetings with the managers involved with emergency preparedness training and legal issues of the facility, the following levels of decision errors were determined to be acceptable: the probability of deciding that a building is emergency prepared should not exceed 1% when in fact 30% or more of its occupants are DUDs; also, when no more than 10% of the occupants of a building are DUDs, the probability that the building will be declared emergency prepared is at least 90%. So the required Decision Error II rate is no more than 1%, and the required Decision Error I rate is no more than 10%. Note that the decision error rates must be specified relative to the true percentage of DUDs in a building. Figure 2.8 illustrates these performance goals for the sampling design (see USEPA, 1994a).

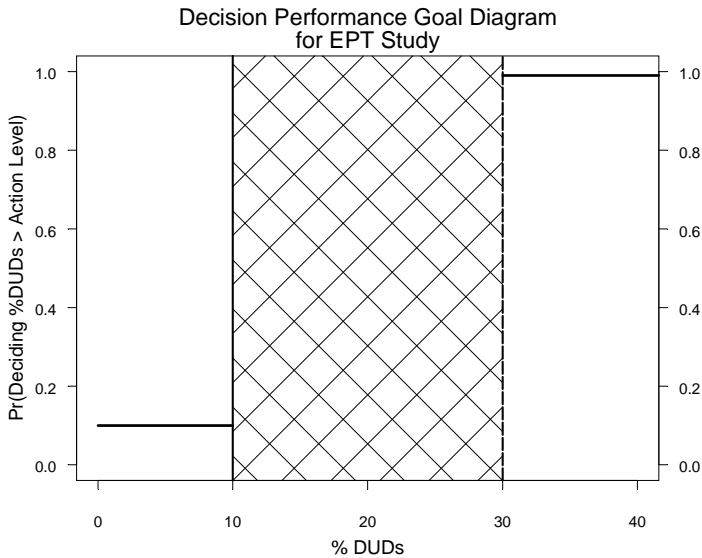


Figure 2.8 Decision performance goal diagram for the emergency preparedness training study with the null hypothesis that the %DUDs is less than the action level of 30%

7. Optimize the Design

A representative sample from each building will be taken using simple random sampling without replacement. The probability calculation related to this method of sampling is based on the hypergeometric distribution (explained in detail in Chapter 4). Table 2.9 lists the number of employees residing in 3 of the 20 buildings, along with the required sample size for each building and the maximum number of DUDs that can be present in the sample without classifying the building as being unprepared for an emergency. The probabilities of making the two types of decision errors are also given in the table.

Building	# of Occupants	Sample Size	Maximum # DUDs (ϵ)	Pr(Error I) (in %)	Pr(Error II) (in %)
A	160	41	6	7.8	0.9
D	247	46	7	6.7	1
F	342	47	7	7.5	0.9

Table 2.9 Sampling plans with probability of Decision Error I $\leq 10\%$ and probability of Decision Error II $\leq 1\%$.

For example, Building A has 160 occupants. A random sample of 41 occupants from the building must be interviewed and if no more than 6 of those interviewed do not know EP facts 1 to 4, then the building is declared emergency prepared. The probability of facing a compliance issue by failing to identify Building A as unprepared for building emergencies when in fact 30% or more of its occupants are DUDs is 0.9% (probability of Error II). On the other hand, the probability of wrongly identifying Building A as unprepared for emergencies and wasting money on training when in fact no more than 10% of the occupants are DUDs is 7.8% (probability of Error I).

Figure 2.9 displays the power curve for the sampling design for Building A drawn on top of the decision performance goal diagram shown in Figure 2.8. Here you can see that this particular sampling design satisfies the performance goals.

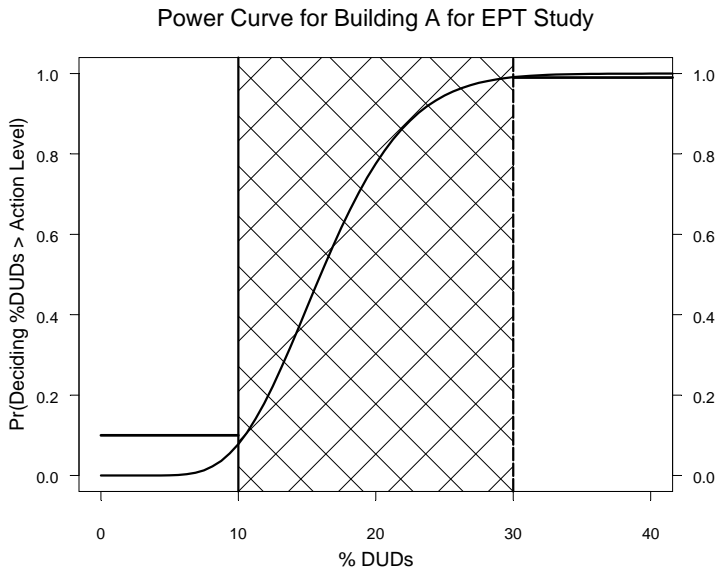


Figure 2.9 Power curve for Building A with the null hypothesis that the %DUDs is less than the action level of 30%, using a sample size of $n = 41$ and maximum number of DUDs of $t = 6$

To explore how the required sample size is affected by our specification of the Decision Error II rate, we create Table 2.10, which is similar to Table 2.9 except that we allow the Decision Error II rate to be 5% instead of 1%. Comparing the required sample sizes in these two tables, you can see that about 10 to 15 more people per building have to be interviewed to reduce the Decision Error II rate from 5% to 1%.

Building	# of Occupants	Sample Size	Maximum # DUDs (ϵ)	Pr(Error I) (in %)	Pr(Error II) (in %)
A	160	31	5	6	4
D	247	32	5	8	4
F	342	32	5	8	4

Table 2.10 Sampling plans with probability of Decision Error I $\leq 10\%$ and probability of Decision Error II $\leq 5\%$.

In the discussion above and in the recommended sampling plans, we have set the highest acceptable percentage of DUDs in a building at 30%. If this percentage is lowered, then a larger sample size is required to ensure the same decision error rates of 1% or 5% for Decision Error II and 10% for Decision Error I. Table 2.11 gives the required sample sizes for the case when the passing percentage is “no more than 20% DUDS,” the Decision Error II rate is 1%, and the Decision Error I rate is 10%. Table 2.12 shows the same thing, except that the Decision Error II rate is allowed to be 5% instead of 1%. The required sample size for each building in Table 2.11 is nearly two times the corresponding sample size given in Table 2.9.

Building	# of Occupants	Sample Size	Maximum # DUDs (ϵ)	Pr(Error I) (in %)	Pr(Error II) (in %)
A	160	87	11	7	1
D	247	104	13	10	1
F	342	119	15	8	0.9

Table 2.11 Sampling plans with probability of Decision Error I $\leq 10\%$ and probability of Decision Error II $\leq 1\%$, with maximum percentage of DUDs set to 20%

Building	# of Occupants	Sample Size	Maximum # DUDs (ϵ)	Pr(Error I) (in %)	Pr(Error II) (in %)
A	160	68	9	8	5
D	247	83	11	9	4
F	342	84	11	10	5

Table 2.12 Sampling plans with probability of Decision Error I $\leq 10\%$ and probability of Decision Error II $\leq 5\%$, with maximum percentage of DUDs set to 20%

Figure 2.10 illustrates how the sampling rate (percentage of occupants selected to be interviewed) increases as the maximum allowed percentage of DUDs decreases from 50% to 10% for Buildings D and F with 247 and 342 occupants, respectively. The figure shows a slow increase followed by a rapid increase. We see that the curves start an uphill climb at around 30%.

This documents and, we believe, also justifies the choice of 30% as the highest acceptable level of DUDs.

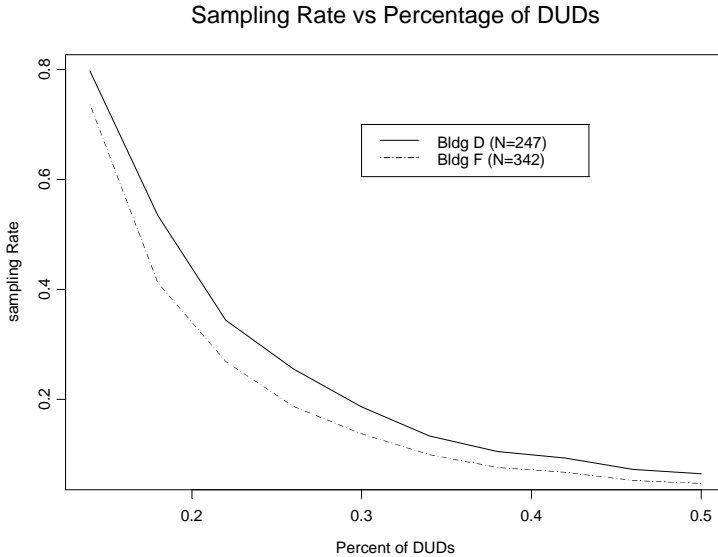


Figure 2.10 Sampling rate vs. maximum allowed percentage of DUDs for Buildings D and F

SUMMARY

- The first and most important step of any environmental study is to define the objectives of the study and design the sampling program.
- The *scientific method* recognizes the fact that our environment is constantly changing and that any of these changes may create an observed effect which may or may not be related to some cause we have hypothesized.
- The basic scientific method consists of forming a hypothesis, performing an experiment using an experimental (exposed) group and a control group, analyzing the results of the experiment, and revising the hypothesis.
- The *Data Quality Objectives (DQO) process* is a formalization of the scientific method that is a systematic way to create a legitimate and effective sampling and analysis plan.
- The term *population* is defined operationally by the question we ask: it is the entire collection of measurements about which we want to make a statement.

- A **sample** is defined as some subset of a population. The statistical definition of the word sample (a selection of individual population members) should not be confused with the more common meaning of a **physical sample** of soil (e.g., 10g of soil), water (e.g., 5ml of water), air (e.g., 20 cc of air), etc.
- **Probability sampling** or **random sampling** involves using a random mechanism to select samples from the population. All statistical methods used to quantify uncertainty assume some form of random sampling has been used to obtain a sample.
- Decisions involving the environment can often be put into the **hypothesis testing framework**. You choose a null and alternative hypothesis, then decide which one is probably true based on the information you have. You then make a decision based on your belief.
- Common mistakes in environmental studies include lack of samples from proper control populations, using judgment sampling instead of random sampling, failing to randomize over potentially influential factors, and collecting too few samples to have a high degree of confidence in the results.
- The DQO process consists of seven steps: state the problem, identify the decision, identify inputs to the decision, define the study boundaries, develop a decision rule, specify acceptable limits on decision errors, and optimize the design.
- The DQO process should include developing a **quality assurance project plan (QAPP)** to ensure the integrity of the data collected for the study.
- Optimizing a sampling plan requires knowledge of the various sources of variability. It is usually a good idea to perform a pilot study to estimate the magnitudes of the sources of variability, and to “fine-tune” the QA/QC procedures.
- This chapter describes four methods of random sampling: simple random sampling (SRS), stratified random sampling, composite sampling, and ranked set sampling (RSS). Other methods of random sampling include two-stage and multi-stage random sampling, double random sampling, sequential random sampling, and adaptive random sampling.

EXERCISES

- 2.1. Concrete pipes that have been used to transport crude oil from the field to a refinery are passing in the proximity of a town. There are complaints that the pipes have leaked. A judgment sampling plan (sample the joints in the pipeline) is being recommended as a preliminary step in investigating the complaints. How would you proceed? Discuss the pros and cons of judgment sampling in this context.
- 2.2. Soil excavated from a contaminated site has been placed in 200-gallon barrels and stored at a temporary storage facility. Barrels containing very high contamination are to be disposed of at a permanent disposal site (burial ground), and the rest of the barrels will be allowed to stay in the temporary facility. The analysis costs are very high and therefore disposal by batches of five barrels is being contemplated. Explain how you would use compositing in this project to save analysis costs.
- 2.3. Suppose for the elephant example discussed in the section on ranked set sampling that the proportion of calves is 25% and the proportion of mothers is 75%. Compute the probability that the ranked set sample consists of only calves. Compute the probability of obtaining such a sample under SRS and show that RSS gives rise to such extreme samples less often.
- 2.4. In the elephant example we obtained a ranked set sample of size $n = 2$. Extend the example to an RSS of size $n = 3$ and show that RSS is superior to SRS in that it gives rise to extreme samples less often.
- 2.5. For the case study discussed at the end of this chapter, the number of DUDs in a sample selected from a specific building is a hypergeometric random variable. Use the S-PLUS menu or the S-PLUS function `phyper` to verify the decision error rates shown in Table 2.9. (See Chapter 4 for an explanation of the hypergeometric distribution.)

3 LOOKING AT DATA

What is Going On?

Once you have a collection of observations from your environmental study, you should thoroughly examine the data in as many ways as possible and relevant. When the first widely available commercial statistical software packages came out in the 1960s, the emphasis was on statistical summaries of data, such as means, standard deviations, and measures of skew and kurtosis. It is still true that “a picture is worth a thousand words,” and no amount of summary or descriptive statistics can replace a good graph to explain your data. John Tukey coined the acronym **EDA**, which stands for ***Exploratory Data Analysis***. Helsel and Hirsch (1992, Chapters 1, 2, and 16) and USEPA (1996) give a good overview of statistical and graphical methods for exploring environmental data. Cleveland (1993, 1994) and Chambers et al. (1983) are excellent general references for methods of graphing data. This chapter discusses the use of summary statistics and graphs to describe and look at environmental data.

SUMMARY STATISTICS

Summary statistics (also called *descriptive statistics*) are numbers that you can use to summarize the information contained in a collection of observations. Summary statistics are also called *sample statistics* because they are statistics computed from a sample; they do not describe the whole population.

One way to classify summary or descriptive statistics is by what they measure: location (central tendency), spread (variability), skew (long-tail in one direction), kurtosis (peakedness), etc. Another way to classify summary statistics is by how they behave when unusually extreme observations are present: sensitive vs. robust. Table 3.1 summarizes several kinds of descriptive statistics based on these two classification schemes. In this section we will give an example of computing summary statistics, and then discuss their formulas and what they measure.

Summary Statistics for TcCB Concentrations

The guidance document USEPA (1994b, pp. 6.22–6.25) contains measures of 1,2,3,4-Tetrachlorobenzene (TcCB) concentrations (in parts per billion, usually abbreviated ppb) from soil samples at a “Reference” site and a

“Cleanup” area. The Cleanup area was previously contaminated and we are interested in determining whether the cleanup process has brought the level

Statistic	What It Measures / How It Is Computed	Robust to Extreme Values?
Mean	Center of distribution Sum of observations divided by sample size Where the histogram balances	No
Trimmed Mean	Center of distribution Trim off extreme observations and compute mean Where the trimmed histogram balances	Somewhat, depends on amount of trim
Median	Center of the distribution Middle value or mean of middle values Half of observations are less and half are greater	Very
Geometric Mean	Center of distribution Exponentiated mean of log-transformed observations Estimates true median for a lognormal distribution	Yes
Variance	Spread of distribution Average of squared distances from the mean	No
Standard Deviation	Spread of distribution Square root of variance In same units as original observations	No
Range	Spread of distribution Maximum minus minimum	No
Interquartile Range	Spread of distribution 75 th percentile minus 25 th percentile Range of middle 50% of data	Yes
Median Absolute Deviation	Spread of distribution Median of distances from the median	Yes
Geometric Standard Deviation	Spread of distribution Exponentiated standard deviation of log-transformed observations	No
Coefficient of Variation	Spread of distribution/Center of distribution Standard deviation divided by mean Sometimes multiplied by 100 and expressed as a percentage	No
Skew	How the distribution leans (left, right, or centered) Average of cubed distances from the mean	No
Kurtosis	Peakedness of the distribution Average of quartic distances from the mean, then subtract 3	No

Table 3.1 A description of commonly used summary statistics

of TcCB back down to what you would find in soil typical of that particular geographic region. The data are shown in Table 3.2.

Area	Observed TcCB (ppb)						
Reference	0.22	0.23	0.26	0.27	0.28	0.28	0.29
	0.33	0.34	0.35	0.38	0.39	0.39	0.42
	0.42	0.43	0.45	0.46	0.48	0.50	0.50
	0.51	0.52	0.54	0.56	0.56	0.57	0.57
	0.60	0.62	0.63	0.67	0.69	0.72	0.74
	0.76	0.79	0.81	0.82	0.84	0.89	1.11
	1.13	1.14	1.14	1.20	1.33		
Cleanup	ND	0.09	0.09	0.12	0.12	0.14	0.16
	0.17	0.17	0.17	0.18	0.19	0.20	0.20
	0.21	0.21	0.22	0.22	0.22	0.23	0.24
	0.25	0.25	0.25	0.25	0.26	0.28	0.28
	0.29	0.31	0.33	0.33	0.33	0.34	0.37
	0.38	0.39	0.40	0.43	0.43	0.47	0.48
	0.48	0.49	0.51	0.51	0.54	0.60	0.61
	0.62	0.75	0.82	0.85	0.92	0.94	1.05
	1.10	1.10	1.19	1.22	1.33	1.39	1.39
	1.52	1.53	1.73	2.35	2.46	2.59	2.61
	3.06	3.29	5.56	6.61	18.40	51.97	168.64

Table 3.2 TcCB concentrations from USEPA (1994b, pp. 6.22–6.25)

There are 47 observations from the Reference site and 77 in the Cleanup area. Note that in Table 3.2, there is one observation in the Cleanup area coded as “ND,” which stands for nondetect. This means that the concentration of TcCB for this soil sample (if any was present at all) was so small that the procedure used to quantify TcCB concentrations could not reliably measure the true concentration. (We will talk more about why observations are sometimes coded as nondetects in Chapters 9 and 10.) For the purposes of this example, we will assume the nondetect observation is less than the smallest observed value, which is 0.09 ppb, but we will set it to the assumed detection limit of 0.09. (In Chapter 10, we talk extensively about statistical methods for handling data sets containing nondetects.)

Figure 3.1 displays two histograms (see the next section, Graphs for a Single Variable), one for the Reference area TcCB concentrations, and one for the Cleanup area TcCB concentrations. Note that these histograms do not share the same x -axis. Also note that in the histogram for the Cleanup area data, the bars for the three largest observations (18.40, 51.97, and 168.64) do not show up because of the scale of the y -axis. Figure 3.2 displays the same two histograms, but on the (natural) logarithmic scale so that the two histograms can share the same x -axis.

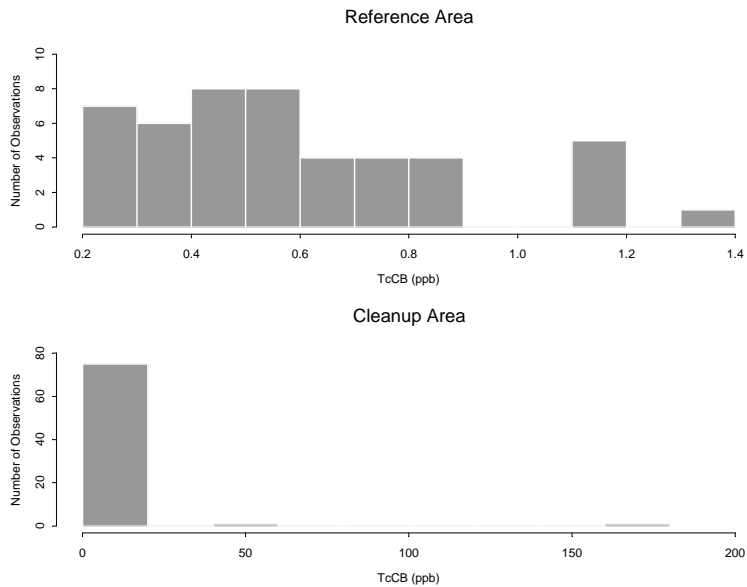


Figure 3.1 Histograms of TcCB concentrations in Reference and Cleanup areas

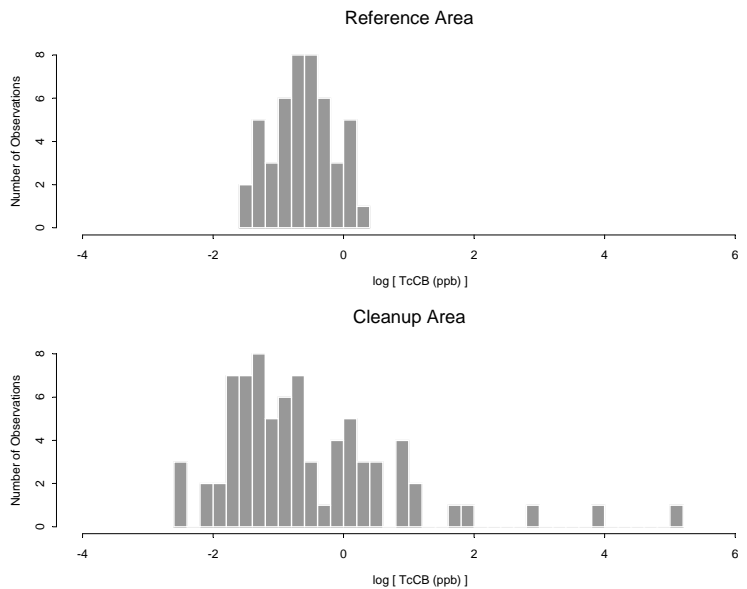


Figure 3.2 Histograms of log-transformed TcCB concentrations in Reference and Cleanup areas

Here are the summary statistics for the original TcCB data.

	Cleanup	Reference
Sample Size:	77	47
# Missing:	0	0
Mean:	3.915	0.5985
Median:	0.43	0.54
10% Trimmed Mean:	0.6846	0.5728
Geometric Mean:	0.5784	0.5382
Skew:	7.566	0.8729
Kurtosis:	61.6	2.993
Min:	0.09	0.22
Max:	168.6	1.33
Range:	168.5	1.11
1st Quartile:	0.23	0.39
3rd Quartile:	1.1	0.75
Standard Deviation:	20.02	0.2836
Geometric Standard Deviation:	3.898	1.597
Interquartile Range:	0.87	0.36
Median Absolute Deviation:	0.3558	0.2669
Coefficient of Variation:	5.112	0.4739

Here are the summary statistics for the log-transformed TcCB data.

	Cleanup	Reference
Sample Size:	77	47
# Missing:	0	0
Mean:	-0.5474	-0.6196
Median:	-0.844	-0.6162
10% Trimmed Mean:	-0.711	-0.6207
Skew:	1.663	0.0307
Kurtosis:	6.889	2.262
Min:	-2.408	-1.514
Max:	5.128	0.2852
Range:	7.536	1.799
1st Quartile:	-1.47	-0.9416
3rd Quartile:	0.09531	-0.2878
Standard Deviation:	1.36	0.468
Interquartile Range:	1.565	0.6538
Median Absolute Deviation:	1.063	0.4825
Coefficient of Variation:	-2.485	-0.7553

Here are a few things to briefly note about the data in Table 3.2 and the summary statistics above:

- Most of the observations in the Cleanup area are similar to or smaller than the observations in the Reference area.
- The Cleanup area data contain several observations that are one, two, and even three orders of magnitude larger than the rest of the observations (this is why we plotted the histograms in Figure 3.2 on a logarithmic scale).

- The mean or average TcCB concentration in the Cleanup area is much larger than in the Reference area (about 4 ppb vs. 0.6 ppb).
- The median TcCB concentration is smaller for the Cleanup area than the Reference area (0.4 ppb vs. 0.5 ppb).
- The standard deviation and coefficient of variation are both two orders of magnitude larger for the Cleanup area.
- The skew of the Cleanup area is an order of magnitude larger than the skew of the Reference area.

All of these characteristics of the Cleanup area data indicate that probably the TcCB contamination has been cleaned up in most of the area, but there are a few places within the Cleanup area with residual contamination. These particular places might be called “hot spots.”

In ENVIRONMENTALSTATS for S-PLUS, the data in Table 3.2 are stored in the data frame `epa.94b.tccb.df` (see the help file `Datasets: USEPA (1994b)`). Here are the steps for using ENVIRONMENTALSTATS for S-PLUS to produce summary statistics for the TcCB data by area.

Menu

To produce summary statistics for the original TcCB data using the ENVIRONMENTALSTATS pull-down menu, follow these steps.

1. Open the Object Explorer, and click on the **Find S-PLUS Objects** button (the binoculars icon).
2. In the Pattern box, type **epa.94b.tccb.df**, then click **OK**.
3. Highlight the shortcut **epa.94b.tccb.df** in the Object column of the Object Explorer.
4. On the S-PLUS menu bar, make the following menu choices: **EnvironmentalStats>EDA>Summary Statistics**. This will bring up the Full Summary Statistics dialog box.
5. In the Data Set box, make sure **epa.94b.tccb.df** is selected.
6. In the Variable(s) box, choose **TcCB**.
7. In the Grouping Variables box, select **Area**.
8. Click **OK** or **Apply**.

To produce summary statistics for the log-transformed TcCB data using the ENVIRONMENTALSTATS pull-down menu, it will simplify things if we first make a new data frame called `new.epa.94b.tccb.df` to contain the original data and the log-transformed TcCB observations. To create the data frame `new.epa.94b.tccb.df`, follow these steps.

1. Highlight the shortcut **epa.94b.tccb.df** in the Object column of the Object Explorer.
2. On the S-PLUS menu bar, make the following menu choices: **Data>Transform**. This will bring up the Transform dialog box.

3. In the Target Column box, type **log.TcCB**.
4. In the Variable box, choose **TcCB**.
5. In the Function box choose **log**.
6. Click on the **Add** button, then click **OK**. At this point, you will get a warning message telling you that you have created a new copy of the data frame `epa.94b.tccb.df` that masks the original copy. Close the message window. Also, the modified data frame pops up in a data window. Close the data window.
7. In the left-hand column of the Object Explorer, click on the **Data** folder. In the right-hand column of the Object Explorer, right-click on **epa.94b.tccb.df** and choose **Properties**. In the Name box rename this data frame to **new.epa.94b.tccb.df** and click **OK**.

To produce summary statistics for the log-transformed TcCB data using the ENVIRONMENTALSTATS pull-down menu, follow these steps.

1. In the Object Explorer, highlight the shortcut **new.epa.94b.tccb.df** in the Object column (right-hand column).
2. On the S-PLUS menu bar, make the following menu choices: **EnvironmentalStats>EDA>Summary Statistics**. This will bring up the Full Summary Statistics dialog box.
3. In the Data Set box, make sure **new.epa.94b.tccb.df** is selected.
4. In the Variable(s) box, choose **log.TcCB**.
5. In the Grouping Variables box, select **Area**.
6. Click on the **Statistics** tab and deselect (uncheck) **Geometric Mean** and **Geometric Standard Deviation**. Click **OK** or **Apply**.

Command

To produce summary statistics for the original TcCB data using the S-PLUS Command or Script Window, type these commands.

```
attach(epa.94b.tccb.df)
full.summary(split(TcCB, Area))
```

To produce summary statistics for the log-transformed TcCB data, type this command.

```
full.summary(split(log(TcCB), Area))
detach()
```

Formulas for Summary Statistics

The formulas for various kinds of summary statistics are shown below. In all of these formulas and throughout this book, we will denote the n observations by:

$$x_1, x_2, \dots, x_n$$

and denote the observations ordered from smallest to largest by:

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}$$

That is, $x_{(1)}$ denotes the smallest value and $x_{(n)}$ denotes the largest value.

Measures of Location (Central Tendency)

Equations (3.1) to (3.4) below show the formulas for four measures of location or central tendency. Often in environmental statistics, we are very interested in the central tendency of a group of measures, either because we want to compare the central tendency to some standard, or because we would like to compare the central tendency of data from one area with the central tendency of data from another area.

Mean

The *mean* (sometimes called *average*) is simply the sum of the observations divided by the sample size.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.1)$$

It indicates approximately where the histogram balances (it is not necessarily exactly where the histogram balances because of the subjectivity of choosing the histogram classes). For example, looking at Figure 3.1 and the summary statistics above, we see that the histogram for the Reference area TcCB data balances at about 0.6 ppb, whereas the histogram for the Cleanup area balances at about 4 ppb. These differences in means between the two areas also demonstrate that the mean is sensitive to extreme values.

Trimmed Mean

The *trimmed mean* involves first trimming off a certain percentage of the smallest and largest observations, and then taking the mean of what is left (Helsel and Hirsch, 1992, p. 7; Hoaglin et al., 1983, pp. 306–311). In the formula below, α is some number between 0 and 0.5 that denotes the trimming fraction, and $[y]$ denotes the largest integer less than or equal to y .

$$\bar{x}_{trimmed} = \frac{1}{n - 2[\alpha n]} \sum_{i=[\alpha n]+1}^{n-[\alpha n]} x_{(i)} \quad (3.2)$$

Because we purposely trim off extreme observations, the trimmed mean is not as sensitive to extreme values as the mean. For the TcCB data, you can see that the mean for the Cleanup area is about 4 ppb, but the 10% trimmed mean is only about 0.7 ppb (very close to the mean for the Reference area).

Median

The **median** is simply the 50% trimmed mean (Helsel and Hirsch, 1992, pp. 5–6; Hoaglin et al., 1983, p. 308). That is, if there is an odd number of observations, the median is the middle value, and if there is an even number of observations, the median is the mean of the two middle values.

$$Median = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{if } n \text{ is even} \end{cases} \quad (3.3)$$

Half of the observations lie below the median and half of them lie above the median.

The median is very robust to extreme values. For example, although the mean TcCB concentration in the Cleanup area is much larger than in the Reference area (about 4 ppb vs. 0.6 ppb), the median TcCB concentration is smaller for the Cleanup area than the Reference area (0.4 ppb vs. 0.5 ppb). You could take almost half of your data and keep increasing it and the median would stay the same.

Geometric Mean

The **geometric mean** is often used to describe positive-valued data. It is the exponentiated mean of the log-transformed observations (Helsel and Hirsch, 1992, p. 6). That is, you take the logarithms of the original observations, compute the mean of these transformed observations, then exponentiate this mean:

$$\bar{x}_g = \exp \left[\frac{1}{n} \sum_{i=1}^n \log(x_i) \right] \quad (3.4)$$

The geometric mean estimates the true *median* for a lognormal distribution (see Chapter 4 for an explanation of the lognormal distribution). For example, note that for the Reference area TcCB data, both the median and geometric mean are 0.54 ppb. The geometric mean is always less than or equal to the sample mean, with equality only if all the observations are the same value (Zar, 1999, p. 28).

Just as the median is robust to extreme observations, so is the geometric mean. The geometric mean for the Cleanup area data is about the same as for the Reference area (0.6 vs. 0.5 ppb).

Measures of Spread (Variability)

Equations (3.5) to (3.16) below show the formulas for seven measures of spread or variability. The spread of a distribution lets us know how well we can characterize it. If the spread is relatively small, then the sample mean or median is a fairly “representative” observation, whereas if the spread is large, then several observations could be much smaller or much larger than the sample mean or median. When we are comparing chemical concentrations from two or more areas, it is also useful to see whether the variability in the data is about the same for all of the areas. If it is not, this may indicate that something unusual is going on. For example, looking at the log-transformed TcCB data in Figure 3.2 and the associated summary statistics, we can see that the central tendency (mean, median, etc.) of the Reference and Cleanup area is about the same, but the spread is much larger in the Cleanup area.

Range

Probably the simplest measure of spread is the *range* of the data; the distance between the largest and smallest value.

$$Range = x_{(n)} - x_{(1)} \quad (3.5)$$

The range quickly gives you an idea about the differences in the orders of magnitude of the observations. For the Reference area TcCB data, the range is about 1 ppb, whereas it is about 169 ppb for the Cleanup area. Obviously, the range is very sensitive to extreme observations.

Interquartile Range

The **interquartile range** (often abbreviated **IQR**) is a modified form of the range. The 25th, 50th, and 75th percentiles of the data are also called the quartiles of the data, so the interquartile range is the distance between the 75th percentile of the data and the 25th percentile (Chambers et al., 1983, p. 21; Helsel and Hirsch, 1992, p. 8; Hoaglin et al., 1983, pp. 38, 59).

$$IQR = x_{0.75} - x_{0.25} \quad (3.6)$$

In the above equation, x_p denotes the $p100^{\text{th}}$ percentile of the data. In Chapter 5 we will talk about how to compute the percentiles for a set of observations. For now all you need to know is that for the $p100^{\text{th}}$ percentile, about $p100\%$ of the observations are less than this number and about $(1-p)100\%$ of the observations are greater than this number.

Unlike the range, the interquartile range is not affected by a few extreme observations; it measures only the range of the middle 50% of the data. For the TcCB data the IQR is about 0.4 ppb for the Reference area and 0.9 ppb for the Cleanup area.

Variance

The **variance** is the mean or average of the squared distances between each observation and the mean.

$$s_{mm}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.7)$$

The sample variance estimates the population variance (see Chapter 5). The formula above is called the method of moments estimator (see Chapter 5), but the more commonly used formula for the sample variance is the unbiased estimator.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.8)$$

Equation (3.8) is the one used by default in ENVIRONMENTALSTATS for S-PLUS and S-PLUS.

For reasons related to estimation (see Chapter 5) and hypothesis testing (see Chapter 7), the variance and standard deviation (see below) are the two

most commonly used statistics to quantify the spread of a set of observations. Unlike the interquartile range, the variance is very sensitive to extreme values, even more so than the mean, because it involves squaring the distances between the observations and the mean. For the TcCB data, the variance for the Reference area is about 0.08 ppb² whereas for the Cleanup area it is about 400 ppb².

Standard Deviation

The **standard deviation** is simply the square root of the variance. The formula based on the method of moments estimator of variance is:

$$s_{mm} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.9)$$

but the most commonly used formula for the sample standard deviation is based on the unbiased estimator of variance:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.10)$$

Equation (3.10) is the one used for the summary statistics for the TcCB data and is used by default in ENVIRONMENTALSTATS for S-PLUS and S-PLUS.

The standard deviation is often preferred to the variance as a measure of spread because it maintains the original units of the data. Just like the variance, the standard deviation is sensitive to extreme values. For the TcCB data, the standard deviation for the Reference area is about 0.3 ppb whereas for the Cleanup area it is about 20 ppb.

Geometric Standard Deviation

The **geometric standard deviation** is sometimes used to describe positive-valued data (Leidel et al., 1977). It is the exponentiated standard deviation of the log-transformed observations.

$$s_g = e^{s_y} \quad (3.11)$$

where

$$s_Y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (3.12)$$

$$Y_i = \log(x_i) \quad (3.13)$$

Unlike the sample geometric mean, the sample geometric standard deviation does not estimate any population parameter that is usually used to characterize the lognormal distribution.

Median Absolute Deviation

The ***median absolute deviation*** (often abbreviated ***MAD***) is the median of the distances between each observation and the median (Helsel and Hirsch, 1992, pp. 8–9; Hoaglin et al., 1983, pp. 220, 346, 365–368).

$$MAD = \text{Median}(|x_1 - m|, |x_2 - m|, \dots, |x_n - m|) \quad (3.14)$$

where

$$m = \text{Median}(x_1, x_2, \dots, x_n) \quad (3.15)$$

Unlike the variance and standard deviation, the median absolute deviation is unaffected by a few extreme observations. For the TcCB data, the MAD is 0.27 ppb for the Reference area and 0.36 ppb for the Cleanup area. You could take almost half of your data and keep increasing it and the MAD would stay the same.

Coefficient of Variation

The ***coefficient of variation*** (sometimes denoted CV) is simply the ratio of the standard deviation to the mean.

$$CV = \frac{S}{\bar{X}} \quad (3.16)$$

The coefficient of variation is a unitless measure of how spread out the distribution is relative to the size of the mean. It is usually used to characterize positive, right-skewed distributions such as the lognormal distribution (see Chapter 4). It is sometimes multiplied by 100 and expressed as a percentage (Zar, 1999, p. 40). Like the mean and standard deviation, the coefficient of variation is sensitive to extreme values. For the TcCB data, the CV is 0.5 for the Reference area and ten times as large for the Cleanup area.

Measures of Deviation from a Symmetric or Bell-Shaped Distribution

Equations (3.17) to (3.19) below show the formulas for two statistics that are used to measure deviation from a symmetric or bell-shaped histogram: the skew and kurtosis. A bell-shaped (and thus symmetric) histogram is a good indication that the data may be modeled with a normal (Gaussian) distribution (see Chapter 4). Many statistical hypothesis tests assume the data follow a normal distribution (see Chapter 7). In the days before it was easy to create plots and perform goodness-of-fit tests with computer software, the skew and kurtosis were often reported. Nowadays, they are not so widely reported, although they are still used to fit distributions in the system of Pearson curves (Johnson et al., 1994, pp. 15–25).

Skew

The *skew* or *coefficient of skewness* is based on the mean or average of the cubed distances between each observation and the mean. The average of the cubed distances is divided by the cube of the standard deviation to produce a unitless measure.

$$Skew_{mm} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s_{mm}^3} \quad (3.17)$$

The formula above uses method of moments estimators. This is the formula that is used by default to compute the skew in ENVIRONMENTALSTATS for S-PLUS. Another formula is sometimes used based on unbiased estimators.

$$Skew = \frac{\frac{n}{(n-1)(n-2)} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3} \quad (3.18)$$

The skew measures how the observations are distributed about the mean. If the histogram is fairly symmetric, then the skew is 0 or close to 0. If there are a few or several large values to the right of the mean (greater than the mean) but not to the left of the mean, the skew is positive and the histogram is said to be **right skewed** or **positively skewed**. If there are a few or several small values to the left of the mean (less than the mean) but not to the right of the mean, the skew is negative and the histogram is said to be **left skewed** or **negatively skewed**.

Because environmental data usually involve measures of chemical concentrations, and concentrations cannot fall below 0, environmental data often tend to be positively skewed (see Figure 3.1). For the log-transformed TcCB data shown in Figure 3.2, the Reference area has a skew of about 0.03 since this histogram is close to being symmetric, but the Cleanup area has a skew of about 1.7.

Kurtosis

The **kurtosis** or **coefficient of kurtosis** is based on the average of the distances between each observation and the mean raised to the 4th power. The average of these distances raised to the 4th power is divided by the square of the standard deviation to produce a unitless measure. The formula below uses method of moments estimators. This is the formula that is used by default to compute the kurtosis in ENVIRONMENTALSTATS for S-PLUS.

$$Kurtosis_{mm} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s_{mm}^4} \quad (3.19)$$

The kurtosis measures how peaked the histogram is relative to an idealized bell-shaped histogram. This idealized bell-shaped histogram is based on the normal (Gaussian) distribution (see Chapter 4), which has a kurtosis of 3. If the histogram has too many observations in the tails compared to the idealized histogram then the kurtosis is larger than 3. If the histogram has short tails and most of the observations are tightly clustered around the mean, then the kurtosis is less than 3. For the log-transformed TcCB data shown in Figure 3.2, the kurtosis for the Reference area is about 2, whereas it is about 7 for the Cleanup area.

GRAPHS FOR A SINGLE VARIABLE

One of the main strengths of S-PLUS and its add-on modules ENVIRONMENTALSTATS for S-PLUS and S+SPATIALSTATS is the great variety