Techniques of Water-Resources Investigations of the United States Geological Survey

Book 4, Hydrologic Analysis and Interpretation

Chapter A3

Statistical Methods in Water Resources

By D.R. Helsel and R.M. Hirsch

SUPERSEDED

Superseded by U.S. Geological Survey Techniques and Methods, book 4, chap. A3
https://doi.org/10.3133/tm4A3

U.S. DEPARTMENT OF THE INTERIOR
GALE A. NORTON, Secretary

U.S. GEOLOGICAL SURVEY
Charles G. Groat, Director

September 2002

Superseded by U.S. Geological Survey Techniques and Methods, book 4, chap. A3

The use of firm, trade, and brand names in this report is for identification purposes only and does not constitute endorsement by the U.S. Geological Survey.

Publication available at:
http://water.usgs.gov/pubs/twri/twri4a3/

Table of Contents

Preface xi

Chapter 1 Summarizing Data 1

1.1 Characteristics of Water Resources Data 2

1.2 Measures of Location 3

1.2.1 Classical Measure -- the Mean 3

1.2.2 Resistant Measure -- the Median 5

1.2.3 Other Measures of Location 6

1.3 Measures of Spread 7

1.3.1 Classical Measures 7

1.3.2 Resistant Measures 8

1.4 Measures of Skewness 9

1.4.1 Classical Measure of Skewness 9

1.4.2 Resistant Measure of Skewness 10

1.5 Other Resistant Measures 10

1.6 Outliers 11

1.7 Transformations 12

1.7.1 The Ladder of Powers 12

Chapter 2 Graphical Data Analysis 17

2.1 Graphical Analysis of Single Data Sets 19

2.1.1 Histograms 19

2.1.2 Stem and Leaf Diagrams 20

2.1.3 Quantile Plots 22

2.1.4 Boxplots 24

2.1.5 Probability Plots 26

2.2 Graphical Comparisons of Two or More Data Sets 35

2.2.1 Histograms 35

2.2.2 Dot and Line Plots of Means, Standard Deviations 35

2.2.3 Boxplots 38

2.2.4 Probability Plots 40

2.2.5 Q-Q Plots 41

2.3 Scatterplots and Enhancements 45

2.3.1 Evaluating Linearity 45
2.3.2 Evaluating Differences in Location on a Scatterplot 47
2.3.3 Evaluating Differences in Spread 50
2.4 Graphs for Multivariate Data 51
2.4.1 Profile Plots 51
2.4.2 Star Plots 53
2.4.3 Trilinear Diagrams 56
2.4.4 Plots of Principal Components 58
2.4.5 Other Multivariate Plots 59

Chapter 3 Describing Uncertainty 65
3.1 Definition of Interval Estimates 66
3.2 Interpretation of Interval Estimates 67
3.3 Confidence Intervals for the Median 70
3.3.1 Nonparametric Interval Estimate for the Median 70
3.3.2 Parametric Interval Estimate for the Median 73
3.4 Confidence Intervals for the Mean 74
3.4.1 Symmetric Confidence Interval for the Mean 75
3.4.2 Asymmetric Confidence Interval for the Mean 76
3.5. Nonparametric Prediction Intervals 76
3.5.1 Two-Sided Nonparametric Prediction Interval 77
3.5.2 One-Sided Nonparametric Prediction Interval 78
3.6 Parametric Prediction Intervals 80
3.6.1 Symmetric Prediction Interval 80
3.6.2 Asymmetric Prediction Intervals 80
3.7 Confidence Intervals for Percentiles (Tolerance Intervals) 82
3.7.1 Nonparametric Confidence Intervals for Percentiles 83
3.7.2 Nonparametric Tests for Percentiles 84
3.7.3 Parametric Confidence Intervals for Percentiles 88
3.7.4 Parametric Tests for Percentiles 90
3.8 Other Uses for Confidence Intervals 90
3.8.1 Implications of Non-Normality for Detection of Outliers 90
3.8.2 Implications of Non-Normality for Quality Control 91
3.8.3 Implications of Non-Normality for Sampling Design 93

Chapter 4 Hypothesis Tests 97
4.1 Classification of Hypothesis Tests 99
4.1.1 Classification Based on Measurement Scales 99
4.1.2 Classification Based on the Data Distribution 100

4.2 Structure of Hypothesis Tests 101
4.2.1 Choose the Appropriate Test 101
4.2.2 Establish the Null and Alternate Hypotheses 104
4.2.3 Decide on an Acceptable Error Rate α 106
4.2.4 Compute the Test Statistic from the Data 107
4.2.5 Compute the p-Value 108
4.2.6 Make the Decision to Reject H_0 or Not 108
4.3 The Rank-Sum Test as an Example of Hypothesis Testing 109
4.4 Tests for Normality 113

Chapter 5 Differences Between Two Independent Groups 117
5.1 The Rank-Sum Test 118
5.1.1 Null and Alternate Hypotheses 118
5.1.2 Computation of the Exact Test 119
5.1.3 The Large Sample Approximation 121
5.1.4 The Rank Transform Approximation 123
5.2 The t-Test 124
5.2.1 Assumptions of the Test 124
5.2.2 Computation of the t-Test 125
5.2.3 Modification for Unequal Variances 125
5.2.4 Consequences of Violating the t-Test's Assumptions 127
5.3 Graphical Presentation of Results 128
5.3.1 Side-by-Side Boxplots 128
5.3.2 Q-Q Plots 129
5.4 Estimating the Magnitude of Differences Between Two Groups 131
5.4.1 The Hodges-Lehmann Estimator 131
5.4.2 Confidence Interval for $\hat{\Delta}$ 132
5.4.3 Difference Between Mean Values 134
5.4.4 Confidence Interval for $\bar{x}-\bar{y}$ 134

Chapter 6 Matched-Pair Tests 137
6.1 The Sign Test 138
6.1.1 Null and Alternate Hypotheses 138
6.1.2 Computation of the Exact Test 138
6.1.3 The Large Sample Approximation 141
6.2 The Signed-Rank Test 142
6.2.1 Null and Alternate Hypotheses 142
6.2.2 Computation of the Exact Test 143
6.2.3 The Large Sample Approximation 145
6.2.4 The Rank Transform Approximation 147

6.3 The Paired t-Test 147
6.3.1 Assumptions of the Test 147
6.3.2 Computation of the Paired t-Test 148
6.4 Consequences of Violating Test Assumptions 149
6.4.1 Assumption of Normality (t-Test) 149
6.4.2 Assumption of Symmetry (Signed-Rank Test) 150
6.5 Graphical Presentation of Results 150
6.5.1 Boxplots 151
6.5.2 Scatterplots With X=Y Line 151
6.6 Estimating the Magnitude of Differences Between Two Groups 153
6.6.1 The Median Difference (Sign Test) 153
6.6.2 The Hodges-Lehmann Estimator (Signed-Rank Test) 153
6.6.3 Mean Difference (t-Test) 155

Chapter 7 Comparing Several Independent Groups 157
7.1 Tests for Differences Due to One Factor 159
7.1.1 The Kruskal-Wallis Test 159
7.1.2 Analysis of Variance (One Factor) 164
7.2 Tests for the Effects of More Than One Factor 169
7.2.1 Nonparametric Multi-Factor Tests 170
7.2.2 Multi-Factor Analysis of Variance -- Factorial ANOVA 170
7.3 Blocking -- The Extension of Matched-Pair Tests 181
7.3.1 Median Polish 182
7.3.2 The Friedman Test 187
7.3.3 Median Aligned-Ranks ANOVA 191
7.3.4 Parametric Two-Factor ANOVA Without Replication 193
7.4 Multiple Comparison Tests 195
7.4.1 Parametric Multiple Comparisons 196
7.4.2 Nonparametric Multiple Comparisons 200
7.5 Presentation of Results 202
7.5.1 Graphical Comparisons of Several Independent Groups 202
7.5.2 Presentation of Multiple Comparison Tests 205

Chapter 8 Correlation 209
8.1 Characteristics of Correlation Coefficients 210
8.1.1 Monotonic Versus Linear Correlation 210
8.2 Kendall's Tau 212
8.2.1 Computation 212
8.2.2 Large Sample Approximation 213
8.2.3 Correction for Ties 215

8.3 Spearman's Rho 217
8.4 Pearson's r 218

Chapter 9 Simple Linear Regression 221
9.1 The Linear Regression Model 222
9.1.1 Assumptions of Linear Regression 224
9.2 Computations 226
9.2.1 Properties of Least Squares Solutions 227
9.3 Building a Good Regression Model 228
9.4 Hypothesis Testing in Regression 237
9.4.1 Test for Whether the Slope Differs from Zero 237
9.4.2 Test for Whether the Intercept Differs from Zero 238
9.4.3 Confidence Intervals on Parameters 239
9.4.4 Confidence Intervals for the Mean Response 240
9.4.5 Prediction Intervals for Individual Estimates of y 241
9.5 Regression Diagnostics 244
9.5.1 Measures of Outliers in the x Direction 246
9.5.2 Measures of Outliers in the y Direction 246
9.5.3 Measures of Influence 248
9.5.4 Measures of Serial Correlation 250
9.6 Transformations of the Response (y) Variable 252
9.6.1 To Transform or Not to Transform? 252
9.6.2 Consequences of Transformation of y 253
9.6.3 Computing Predictions of Mass (Load) 255
9.6.4 An Example 257
9.7 Summary Guide to a Good SLR Model 261

Chapter 10 Alternative Methods to Regression 265
10.1 Kendall-Theil Robust Line 266
10.1.1 Computation of the Line 266
10.1.2 Properties of the Estimator 267
10.1.3 Test of Significance 272
10.1.4 Confidence Interval for Theil Slope 273
10.2 Alternative Parametric Linear Equations 274
10.2.1 OLS of X on Y 275
10.2.2 Line of Organic Correlation 276
10.2.3 Least Normal Squares 278
10.2.4 Summary of the Applicability of OLS, LOC and LNS 280
10.3 Weighted Least Squares 280
10.4 Iteratively Weighted Least Squares 283

10.5 Smoothing 285
10.5.1 Moving Median Smooths 285
10.5.2 LOWESS 287
10.5.3 Polar Smoothing 291

Chapter 11 Multiple Linear Regression 295

11.1 Why Use MLR? 296
11.2 MLR Model 296
11.3 Hypothesis Tests for Multiple Regression 297
11.3.1 Nested F Tests 297
11.3.2 Overall F Test 298
11.3.3 Partial F Tests 298
11.4 Confidence Intervals 299
11.4.1 Variance-Covariance Matrix 299
11.4.2 Confidence Intervals for Slope Coefficients 299
11.4.3 Confidence Intervals for the Mean Response 300
11.4.4 Prediction Intervals for an Individual y 300
11.5 Regression Diagnostics 300
11.5.1 Partial Residual Plots 301
11.5.2 Leverage and Influence 301
11.5.3 Multi-Collinearity 305
11.6 Choosing the Best MLR Model 309
11.6.1 Stepwise Procedures 310
11.6.2 Overall Measures of Quality 313
11.7 Summary of Model Selection Criteria 315
11.8 Analysis of Covariance 316
11.8.1 Use of One Binary Variable 316
11.8.2 Multiple Binary Variables 318

Chapter 12 Trend Analysis 323

12.1 General Structure of Trend Tests 324
12.1.1 Purpose of Trend Testing 324
12.1.2 Approaches to Trend Testing 325
12.2 Trend Tests With No Exogenous Variable 326
12.2.1 Nonparametric Mann-Kendall Test 326
12.2.2 Parametric Regression of Y on T 328
12.2.3 Comparison of Simple Tests for Trend 328
12.3 Accounting for Exogenous Variables 329
12.3.1 Nonparametric Approach 334
12.3.2 Mixed Approach 335

12.3.3 Parametric Approach 335
12.3.4 Comparison of Approaches 336
12.4 Dealing With Seasonality 337
12.4.1 The Seasonal Kendall Test 338
12.4.2 Mixture Methods 340
12.4.3 Multiple Regression With Periodic Functions 341
12.4.4 Comparison of Methods 342
12.4.5 Presenting Seasonal Effects 343
12.4.6 Differences Between Seasonal Patterns 344
12.5 Use of Transformations in Trend Studies 346
12.6 Monotonic Trend versus Two Sample (Step) Trend 348
12.7 Applicability of Trend Tests With Censored Data 352

Chapter 13 Methods for Data Below the Reporting Limit 357
13.1 Methods for Estimating Summary Statistics 358
13.1.1 Simple Substitution Methods 358
13.1.2 Distributional Methods 360
13.1.3 Robust Methods 362
13.1.4 Recommendations 362
13.1.5 Multiple Reporting Limits 364
13.2 Methods for Hypothesis Testing 366
13.2.1 Simple Substitution Methods 366
13.2.2 Distributional Test Procedures 367
13.2.3 Nonparametric Tests 367
13.2.4 Hypothesis Testing With Multiple Reporting Limits 369
13.2.5 Recommendations 370
13.3 Methods For Regression With Censored Data 371
13.3.1 Kendall's Robust Line Fit 371
13.3.2 Tobit Regression 371
13.3.3 Logistic Regression 372
13.3.4 Contingency Tables 373
13.3.5 Rank Correlation Coefficients 373
13.3.6 Recommendations 374

Chapter 14 Discrete Relationships 377
14.1 Recording Categorical Data 378
14.2 Contingency Tables (Both Variables Nominal) 378
14.2.1 Performing the Test for Independence 379
14.2.2 Conditions Necessary for the Test 381
14.2.3 Location of the Differences 382
14.3 Kruskal-Wallis Test for Ordered Categorical Responses 382

14.3.1 Computing the Test 383
14.3.2 Multiple Comparisons 385
14.4 Kendall's Tau for Categorical Data (Both Variables Ordinal) 385
14.4.1 Kendall's τ_b for Tied Data 385
14.4.2 Test of Significance for τ_b 388
14.5 Other Methods for Analysis of Categorical Data 390

Chapter 15 Regression for Discrete Responses 393
15.1 Regression for Binary Response Variables 394
15.1.1 Use of Ordinary Least Squares 394
15.2 Logistic Regression 395
15.2.1 Important Formulae 395
15.2.2 Computation by Maximum Likelihood 396
15.2.3 Hypothesis Tests 397
15.2.4 Amount of Uncertainty Explained, R^2 398
15.2.5 Comparing Non-Nested Models 398
15.3 Alternatives to Logistic Regression 402
15.3.1 Discriminant Function Analysis 402
15.3.2 Rank-Sum Test 402
15.4 Logistic Regression for More Than Two Response Categories 403
15.4.1 Ordered Response Categories 403
15.4.2 Nominal Response Categories 405

Chapter 16 Presentation Graphics 409
16.1 The Value of Presentation Graphics 410
16.2 Precision of Graphs 411
16.2.1 Color 412
16.2.2 Shading 413
16.2.3 Volume and Area 416
16.2.4 Angle and Slope 417
16.2.5 Length 420
16.2 6 Position Along Nonaligned Scales 421
16.2.7 Position Along an Aligned Scale 423
16.3 Misleading Graphics to be Avoided 423
16.3.1 Perspective 423
16.3.2 Graphs With Numbers 426
16.3.3 Hidden Scale Breaks 427
16.3.4 Overlapping Histograms 428

References 433

Appendix A **Construction of Boxplots** **451**

Appendix B **Tables** **456**

Appendix C **Data Sets** **468**

Appendix D **Answers to Exercises** **469**

Index **503**

Preface

This book began as class notes for a course we teach on applied statistical methods to hydrologists of the Water Resources Division, U. S. Geological Survey (USGS). It reflects our attempts to teach statistical methods which are appropriate for analysis of water resources data. As interest in this course has grown outside of the USGS, incentive grew to develop the material into a textbook. The topics covered are those we feel are of greatest usefulness to the practicing water resources scientist. Yet all topics can be directly applied to many other types of environmental data.

This book is not a stand-alone text on statistics, or a text on statistical hydrology. For example, in addition to this material we use a textbook on introductory statistics in the USGS training course. As a consequence, discussions of topics such as probability theory required in a general statistics textbook will not be found here. Derivations of most equations are not presented. Important tables included in all general statistics texts, such as quantiles of the normal distribution, are not found here. Neither are details of how statistical distributions should be fitted to flood data -- these are adequately covered in numerous books on statistical hydrology.

We have instead chosen to emphasize topics not always found in introductory statistics textbooks, and often not adequately covered in statistical textbooks for scientists and engineers. Tables included here, for example, are those found more often in books on nonparametric statistics than in books likely to have been used in college courses for engineers. This book points the environmental and water resources scientist to robust and nonparametric statistics, and to exploratory data analysis. We believe that the characteristics of environmental (and perhaps most other 'real') data drive analysis methods towards use of robust and nonparametric methods.

Exercises are included at the end of chapters. In our course, students compute each type of analysis (t-test, regression, etc.) the first time by hand. We choose the smaller, simpler examples for hand computation. In this way the mechanics of the process are fully understood, and computer software is seen as less mysterious.

We wish to acknowledge and thank several other scientists at the U. S. Geological Survey for contributing ideas to this book. In particular, we thank those who have served as the other instructors at the USGS training course. Ed Gilroy has critiqued and improved much of the material found in this book. Tim Cohn has contributed in several areas, particularly to the sections on bias correction in regression, and methods for data below the reporting limit. Richard Alexander has added to the trend analysis chapter, and Charles Crawford has contributed ideas for regression and ANOVA. Their work has undoubtedly made its way into this book without adequate recognition.

Professor Ken Potter (University of Wisconsin) and Dr. Gary Tasker (USGS) reviewed the manuscript, spending long hours with no reward except the knowledge that they have improved the work of others. For that we are very grateful. We also thank Madeline Sabin, who carefully typed original drafts of the class notes on which the book is based. As always, the responsibility for all errors and slanted thinking are ours alone.

Dennis R. Helsel

Robert M. Hirsch

Reston, VA USA
June, 1991

Citations of trade names in this book are for reference purposes only, and do not reflect endorsement by the authors or by the U. S. Geological Survey

Chapter 1

Summarizing Data

When determining how to appropriately analyze any collection of data, the first consideration must be the characteristics of the data themselves. Little is gained by employing analysis procedures which assume that the data possess characteristics which in fact they do not. The result of such false assumptions may be that the interpretations provided by the analysis are incorrect, or unnecessarily inconclusive. Therefore we begin this book with a discussion of the common characteristics of water resources data. These characteristics will determine the selection of appropriate data analysis procedures.

One of the most frequent tasks when analyzing data is to describe and summarize those data in forms which convey their important characteristics. "What is the sulfate concentration one might expect in rainfall at this location"? "How variable is hydraulic conductivity"? "What is the 100 year flood" (the 99th percentile of annual flood maxima)? Estimation of these and similar summary statistics are basic to understanding data. Characteristics often described include: a measure of the center of the data, a measure of spread or variability, a measure of the symmetry of the data distribution, and perhaps estimates of extremes such as some large or small percentile. This chapter discusses methods for summarizing or describing data.

This first chapter also quickly demonstrates one of the major themes of the book -- the use of robust and resistant techniques. The reasons why one might prefer to use a resistant measure, such as the median, over a more classical measure such as the mean, are explained.

The data about which a statement or summary is to be made are called the **population**, or sometimes the **target population**. These might be concentrations in all waters of an aquifer or stream reach, or all streamflows over some time at a particular site. Rarely are all such data available to the scientist. It may be physically impossible to collect all data of interest (all the water in a stream over the study period), or it may just be financially impossible to collect them. Instead, a subset of the data called the **sample** is selected and measured in such a way that conclusions about the sample may be extended to the entire population. Statistics computed from the sample are only inferences or estimates about characteristics of the population, such as location, spread, and skewness. Measures of location are usually the sample mean and sample median. Measures of spread include the sample standard deviation and sample interquartile range. Use of the term "sample" before each statistic explicitly demonstrates that these only estimate the population value, the population mean or median, etc. As sample estimates are far more common than measures based on the entire population, the term "mean" should be interpreted as the "sample mean", and similarly for other statistics used in this book. When population values are discussed they will be explicitly stated as such.

1.1 Characteristics of Water Resources Data

Data analyzed by the water resources scientist often have the following characteristics:

1. A lower bound of zero. No negative values are possible.
2. Presence of 'outliers', observations considerably higher or lower than most of the data, which infrequently but regularly occur. outliers on the high side are more common in water resources.
3. Positive skewness, due to items 1 and 2. An example of a skewed distribution, the lognormal distribution, is presented in figure 1.1. Values of an observation on the horizontal axis are plotted against the frequency with which that value occurs. These density functions are like histograms of large data sets whose bars become infinitely narrow. Skewness can be expected when outlying values occur in only one direction.
4. Non-normal distribution of data, due to items 1 - 3 above. Figure 1.2 shows an important symmetric distribution, the normal. While many statistical tests assume data follow a normal distribution as in figure 1.2, water resources data often look more like figure 1.1. In addition, symmetry does not guarantee normality. Symmetric data with more observations at both extremes (heavy tails) than occurs for a normal distribution are also non-normal.
5. Data reported only as below or above some threshold (censored data). Examples include concentrations below one or more detection limits, annual flood stages known only to be lower than a level which would have caused a public record of the flood, and hydraulic heads known only to be above the land surface (artesian wells on old maps).
6. Seasonal patterns. Values tend to be higher or lower in certain seasons of the year.

7. Autocorrelation. Consecutive observations tend to be strongly correlated with each other. For the most common kind of autocorrelation in water resources (positive autocorrelation), high values tend to follow high values and low values tend to follow low values.
8. Dependence on other uncontrolled variables. Values strongly covary with water discharge, hydraulic conductivity, sediment grain size, or some other variable.

Methods for analysis of water resources data, whether the simple summarization methods such as those in this chapter, or the more complex procedures of later chapters, should recognize these common characteristics.

1.2 Measures of Location

The mean and median are the two most commonly-used measures of location, though they are not the only measures available. What are the properties of these two measures, and when should one be employed over the other?

1.2.1 Classical Measure -- the Mean

The mean $(\overline{X})$ is computed as the sum of all data values X_i, divided by the sample size n:

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \qquad [1.1]$$

For data which are in one of k groups, equation [1.1] can be rewritten to show that the overall mean depends on the mean for each group, weighted by the number of observations n_i in each group:

$$\overline{X} = \sum_{i=1}^{n} \overline{X}_i \frac{n_i}{n} \qquad [1.2]$$

where $\overline{X}_i$ is the mean for group i. The influence of any one observation X_j on the mean can be seen by placing all but that one observation in one "group", or

$$\overline{X} = \overline{X}_{(j)} \frac{(n-1)}{n} + X_j \bullet \frac{1}{n}.$$

$$= \overline{X}_{(j)} + (X_{(j)} - \overline{X}_{(j)}) \bullet \frac{1}{n}. \qquad [1.3]$$

where $\overline{X}_{(j)}$ is the mean of all observations excluding X_j. Each observation's influence on the overall mean $\overline{X}$ is $(X_j - \overline{X}_{(j)})$, the distance between the observation and the mean excluding that observation. Thus all observations do not have the same influence on the mean. An 'outlier' observation, either high or low, has a much greater influence on the overall mean $\overline{X}$ than does a more 'typical' observation, one closer to its $\overline{X}_{(j)}$.

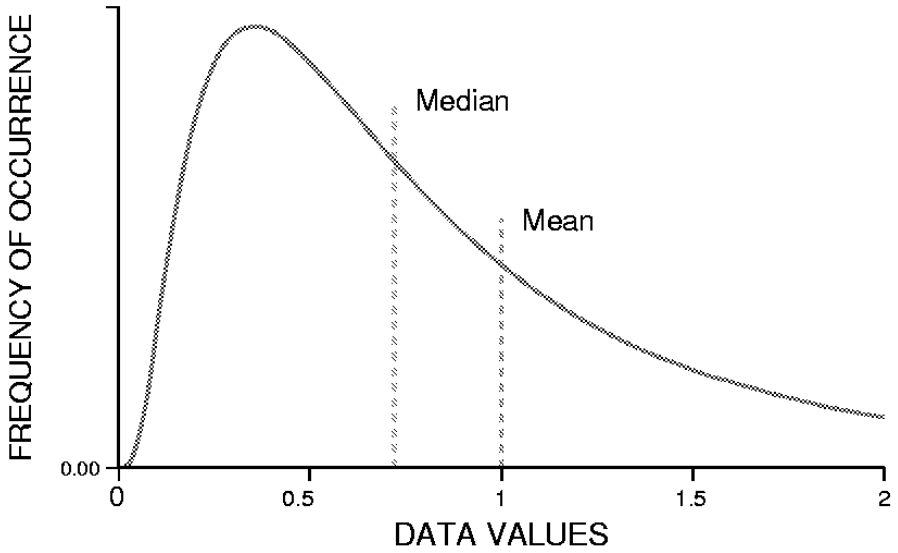


Figure 1.1 Density Function for a Lognormal Distribution

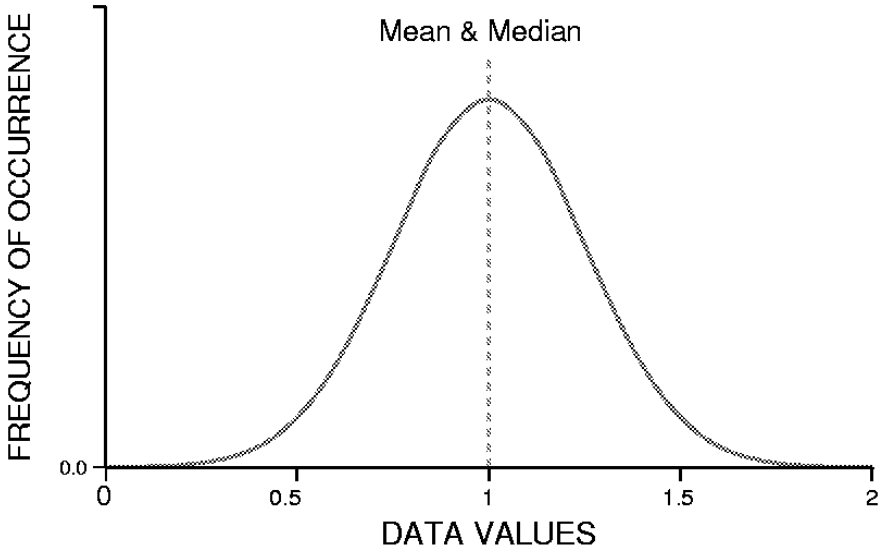


Figure 1.2 Density Function for a Normal Distribution

Another way of illustrating this influence is to realize that the mean is the balance point of the data, when each point is stacked on a number line (figure 1.3a). Data points further from the center exert a stronger downward force than those closer to the center. If one point near the center were removed, the balance point would only need a small adjustment to keep the data set in balance. But if one outlying value were removed, the balance point would shift dramatically (figure 1.3b). This sensitivity to the magnitudes of a small number of points in the data set defines why the mean is not a "resistant" measure of location. It is not resistant to changes in the presence of, or to changes in the magnitudes of, a few outlying observations.

When this strong influence of a few observations is desirable, the mean is an appropriate measure of center. This usually occurs when computing units of mass, such as the average concentration of sediment from several samples in a cross-section. Suppose that sediment concentrations closer to the river banks were much higher than those in the center. Waters represented by a bottle of high concentration would exert more influence (due to greater mass of sediment per volume) on the final concentration than waters of low or average concentration. This is entirely appropriate, as the same would occur if the stream itself were somehow mechanically mixed throughout its cross section.

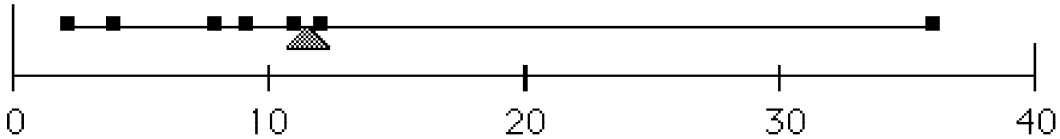


Figure 1.3a The mean (triangle) as balance point of a data set.

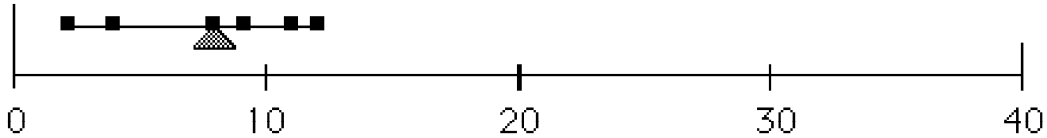


Figure 1.3b Shift of the mean downward after removal of outlier.

1.2.2 Resistant Measure -- the Median

The median, or 50th percentile $P_{0.50}$, is the central value of the distribution when the data are ranked in order of magnitude. For an odd number of observations, the median is the data point which has an equal number of observations both above and below it. For an even number of observations, it is the average of the two central observations. To compute the median, first

rank the observations from smallest to largest, so that x_1 is the smallest observation, up to x_n, the largest observation. Then

$$\text{median } (P_{0.50}) = X_{(n+1)/2} \qquad \text{when n is odd, and}$$

$$\text{median } (P_{0.50}) = \frac{1}{2}(X_{(n/2)} + X_{(n/2)+1}) \qquad \text{when n is even.} \qquad [1.4]$$

The median is only minimally affected by the magnitude of a single observation, being determined solely by the relative order of observations. This resistance to the effect of a change in value or presence of outlying observations is often a desirable property. To demonstrate the resistance of the median, suppose the last value of the following data set (a) of 7 observations were multiplied by 10 to obtain data set (b):

Example 1:

(a)	2 4 8 9 11 11 12	$\bar{X}$ = 8.1	$P_{.50}$= 9
(b)	2 4 8 9 11 11 120	$\bar{X}$ = 23.6	$P_{.50}$= 9

The mean increases from 8.1 to 23.6. The median, the $\frac{(7+1)}{2}$ th or 4th lowest data point, is unaffected by the change.

When a summary value is desired that is not strongly influenced by a few extreme observations, the median is preferable to the mean. One such example is the chemical concentration one might expect to find over many streams in a given region. Using the median, one stream with unusually high concentration has no greater effect on the estimate than one with low concentration. The mean concentration may be pulled towards the outlier, and be higher than concentrations found in most of the streams. Not so for the median.

1.2.3 Other Measures of Location

Three other measures of location are less frequently used: the mode, the geometric mean, and the trimmed mean. The mode is the most frequently observed value. It is the value having the highest bar in a histogram. It is far more applicable for grouped data, data which are recorded only as falling into a finite number of categories, than for continuous data. It is very easy to obtain, but a poor measure of location for continuous data, as its value often depends on the arbitrary grouping of those data.

The geometric mean (GM) is often reported for positively skewed data sets. It is the mean of the logarithms, transformed back to their original units.

$$GM = \exp(\bar{Y}), \qquad \text{where } Y_i = \ln(X_i) \qquad [1.5]$$

(in this book the natural, base e logarithm will be abbreviated **ln**, and its inverse e^x abbreviated **exp(x)**). For positively skewed data the geometric mean is usually quite close to the median. In fact, when the logarithms of the data are symmetric, the geometric mean is an unbiased estimate

of the median. This is because the median and mean logarithms are equal, as in figure 1.2. When transformed back to original units, the geometric mean continues to be an estimate for the median, but is not an estimate for the mean (figure 1.1).

Compromises between the median and mean are available by trimming off several of the lowest and highest observations, and calculating the mean of what is left. Such estimates of location are not influenced by the most extreme (and perhaps anomalous) ends of the sample, as is the mean. Yet they allow the magnitudes of most of the values to affect the estimate, unlike the median. These estimators are called "trimmed means", and any desirable percentage of the data may be trimmed away. The most common trimming is to remove 25 percent of the data on each end -- the resulting mean of the central 50 percent of data is commonly called the "trimmed mean", but is more precisely the 25 percent trimmed mean. A "0% trimmed mean" is the sample mean itself, while trimming all but 1 or 2 central values produces the median. Percentages of trimming should be explicitly stated when used. The trimmed mean is a resistant estimator of location, as it is not strongly influenced by outliers, and works well for a wide variety of distributional shapes (normal, lognormal, etc.). It may be considered a weighted mean, where data beyond the cutoff 'window' are given a weight of 0, and those within the window a weight of 1.0 (see figure 1.4).

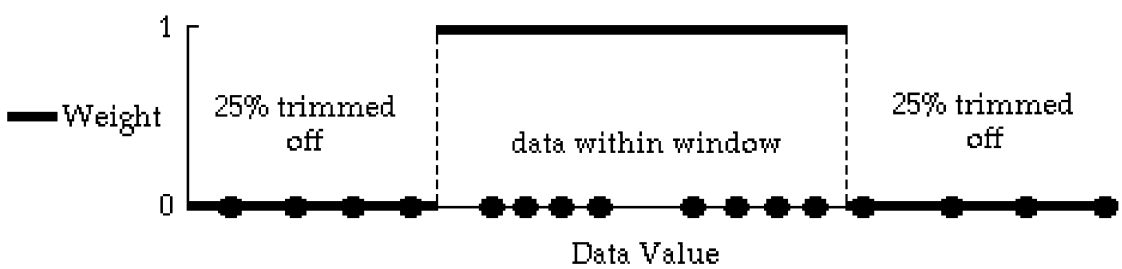


Figure 1.4. Window diagram for the trimmed mean

1.3 Measures of Spread

It is just as important to know how variable the data are as it is to know their general center or location. Variability is quantified by measures of spread.

1.3.1 Classical Measures

The sample variance, and its square root the sample standard deviation, are the classical measures of spread. Like the mean, they are strongly influenced by outlying values.

$$s^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{(n-1)} \qquad \text{sample variance} \qquad [1.6]$$

$$s = \sqrt{s^2} \qquad \text{sample standard deviation} \qquad [1.7]$$

They are computed using the squares of deviations of data from the mean, so that outliers influence their magnitudes even more so than for the mean. When outliers are present these measures are unstable and inflated. They may give the impression of much greater spread than is indicated by the majority of the data set.

1.3.2 Resistant Measures

The interquartile range (IQR) is the most commonly-used resistant measure of spread. It measures the range of the central 50 percent of the data, and is not influenced at all by the 25 percent on either end. It is therefore the width of the non-zero weight window for the trimmed mean of figure 1.4.

The IQR is defined as the 75th percentile minus the 25th percentile. The 75th, 50th (median) and 25th percentiles split the data into four equal-sized quarters. The 75th percentile ($P_{.75}$), also called the upper quartile, is a value which exceeds no more than 75 percent of the data <u>and</u> is exceeded by no more than 25 percent of the data. The 25th percentile ($P_{.25}$) or lower quartile is a value which exceeds no more than 25 percent of the data <u>and</u> is exceeded by no more than 75 percent. Consider a data set ordered from smallest to largest: X_i, i =1,...n. Percentiles (P_j) are computed using equation [1.8]

$$P_j = X_{(n+1)\bullet j} \qquad [1.8]$$

where n is the sample size of X_i, and
j is the fraction of data less than or equal to the percentile value (for the 25th, 50th and 75th percentiles, j= .25, .50, and .75).

Non-integer values of (n+1)•j imply linear interpolation between adjacent values of X. For the example 1 data set given earlier, n=7, and therefore the 25th percentile is $X_{(7+1)\bullet .25}$ or $X_2 = 4$, the second lowest observation. The 75th percentile is X_6 , the 6th lowest observation, or 11. The IQR is therefore 11−4 = 7.

One resistant estimator of spread other than the IQR is the Median Absolute Deviation, or MAD. The MAD is computed by first listing the absolute value of all differences |d| between each observation and the median. The median of these absolute values is then the MAD.

$$\text{MAD}(X_i) = \text{median}\,|d_i|, \qquad \text{where } d_i = X_i - \text{median}(X_i) \qquad [1.9]$$

Comparison of each estimate of spread for the Example 1 data set is as follows. When the last value is changed from 12 to 120, the standard deviation increases from 3.8 to 42.7. The IQR and the MAD remain exactly the same.

data	2	4	8	9	11	11	12	IQR = 11 – 4 = 7
$(X_i - \bar{X})^2$	37.2	16.8	0.01	0.81	8.41	8.41	15.2	$s^2 = (3.8)^2$
$\lvert d_i = X_i - P_{.50} \rvert$	7	5	1	0	2	2	3	MAD=median$\lvert d_i \rvert$=2

data	2	4	8	9	11	11	120	IQR = 11 – 4 = 7
$(X_i - \bar{X})^2$	37.2	16.8	0.01	0.81	8.41	8.41	12,522	$s^2 = (42.7)^2$
$\lvert d_i = X_i - P_{.50} \rvert$	7	5	1	0	2	2	111	MAD=median$\lvert d_i \rvert$=2

1.4 Measures of Skewness

Hydrologic data are typically skewed, meaning that data sets are not symmetric around the mean or median, with extreme values extending out longer in one direction. The density function for a lognormal distribution shown previously as figure 1.1 illustrates this skewness. When extreme values extend the right tail of the distribution, as they do with figure 1.1, the data are said to be skewed to the right, or positively skewed. Left skewness, when the tail extends to the left, is called negative skew.

When data are skewed the mean is not expected to equal the median, but is pulled toward the tail of the distribution. Thus for positive skewness the mean exceeds more than 50 percent of the data, as in figure 1.1. The standard deviation is also inflated by data in the tail. Therefore, tables of summary statistics which include only the mean and standard deviation or variance are of questionable value for water resources data, as those data often have positive skewness. The mean and standard deviation reported may not describe the majority of the data very well. Both will be inflated by outlying observations. Summary tables which include the median and other percentiles have far greater applicability to skewed data. Skewed data also call into question the applicability of hypothesis tests which are based on assumptions that the data have a normal distribution. These tests, called parametric tests, may be of questionable value when applied to water resources data, as the data are often neither normal nor even symmetric. Later chapters will discuss this in much detail, and suggest several solutions.

1.4.1 Classical Measure of Skewness

The coefficient of skewness (g) is the skewness measure used most often. It is the adjusted third moment divided by the cube of the standard deviation:

$$g = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(x_i - \bar{X})^3}{s^3} \qquad [1.10]$$

A right-skewed distribution has positive g; a left-skewed distribution has negative g. Again, the influence of a few outliers is important -- an otherwise symmetric distribution having one outlier will produce a large (and possibly misleading) measure of skewness. For the example 1 data, the g skewness coefficient increases from –0.5 to 2.6 when the last data point is changed from 12 to 120.

1.4.2 Resistant Measure of Skewness

A more resistant measure of skewness is the quartile skew coefficient qs (Kenney and Keeping, 1954):

$$qs = \frac{(P_{.75} - P_{.50}) - (P_{.50} - P_{.25})}{P_{.75} - P_{.25}} \qquad [1.11]$$

the difference in distances of the upper and lower quartiles from the median, divided by the IQR. A right-skewed distribution again has positive qs; a left-skewed distribution has negative qs. Similar to the trimmed mean and IQR, qs uses the central 50 percent of the data. For the example 1 data, qs = (11–9) – (9–4) / (11–4) = –0.43 both before and after alteration of the last data point. Note that this resistance may be a liability if sensitivity to a few observations is important.

1.5 Other Resistant Measures

Other percentiles may be used to produce a series of resistant measures of location, spread and skewness. For example, the 10 percent trimmed mean can be coupled with the range between the 10th and 90th percentiles as a measure of spread, and a corresponding measure of skewness:

$$qs_{.10} = \frac{(P_{.90} - P_{.50}) - (P_{.50} - P_{.10})}{P_{.90} - P_{.10}} \qquad [1.12]$$

to produce a consistent series of resistant statistics. Geologists have used the 16th and 84th percentiles for many years to compute a similar series of robust measures of the distributions of sediment particles (Inman, 1952). However, measures based on quartiles have become generally standard, and other measures should be clearly defined prior to their use. The median, IQR, and quartile skew can be easily summarized graphically using a boxplot (see Chapter 2) and are familiar to most data analysts.

1.6 Outliers

Outliers, observations whose values are quite different than others in the data set, often cause concern or alarm. They should not. They are often dealt with by throwing them away prior to describing data, or prior to some of the hypothesis test procedures of later chapters. Again, they should not. Outliers may be the most important points in the data set, and should be investigated further.

It is said that data on the Antarctic ozone "hole", an area of unusually low ozone concentrations, had been collected for approximately 10 years prior to its actual discovery. However, the automatic data checking routines during data processing included instructions on deleting "outliers". The definition of outliers was based on ozone concentrations found at mid-latitudes. Thus all of this unusual data was never seen or studied for some time. If outliers are deleted, the risk is taken of seeing only what is expected to be seen.

Outliers can have one of three causes:

1. a measurement or recording error.
2. an observation from a population not similar to that of most of the data, such as a flood caused by a dam break rather than by precipitation.
3. a rare event from a single population that is quite skewed.

The graphical methods of the Chapter 2 are very helpful in identifying outliers. Whenever outliers occur, first verify that no copying, decimal point, or other obvious error has been made. If not, it may not be possible to determine if the point is a valid one. The effort put into verification, such as re-running the sample in the laboratory, will depend on the benefit gained versus the cost of verification. Past events may not be able to be duplicated. If no error can be detected and corrected, **outliers should not be discarded based solely on the fact that they appear unusual**. Outliers are often discarded in order to make the data nicely fit a pre-conceived theoretical distribution such as the normal. There is no reason to suppose that they should! The entire data set may arise from a skewed distribution, and taking logarithms or some other transformation may produce quite symmetrical data. Even if no transformation achieves symmetry, outliers need not be discarded. Rather than eliminating actual (and possibly very important) data in order to use analysis procedures requiring symmetry or normality, procedures which are resistant to outliers should instead be employed. If computing a mean appears of little value because of an outlier, the median has been shown to be a more appropriate measure of location for skewed data. If performing a t-test (described later) appears invalidated because of the non-normality of the data set, use a rank-sum test instead.

In short, let the data guide which analysis procedures are employed, rather than altering the data in order to use some procedure having requirements too restrictive for the situation at hand.

1.7 Transformations

Transformations are used for three purposes:

1. to make data more symmetric,
2. to make data more linear, and
3. to make data more constant in variance.

Some water resources scientists fear that by transforming data, results are derived which fit preconceived ideas. Therefore, transformations are methods to 'see what you want to see' about the data. But in reality, serious problems can occur when procedures assuming symmetry, linearity, or homoscedasticity (constant variance) are used on data which do not possess these required characteristics. Transformations can produce these characteristics, and thus the use of transformed variables meets an objective. Employment of a transformation is not merely an arbitrary choice.

One unit of measurement is no more valid a priori than any other. For example, the negative logarithm of hydrogen ion concentration, pH, is as valid a measurement system as hydrogen ion concentration itself. Transformations like the square root of depth to water at a well, or cube root of precipitation volume, should bear no more stigma than does pH. These measurement scales may be more appropriate for data analysis than are the original units. Hoaglin (1988) has written an excellent article on hidden transformations, consistently taken for granted, which are in common use by everyone. Octaves in music are a logarithmic transform of frequency. Each time a piano is played a logarithmic transform is employed! Similarly, the Richter scale for earthquakes, miles per gallon for gasoline consumption, f-stops for camera exposures, etc. all employ transformations. In the science of data analysis, the decision of which measurement scale to use should be determined by the data, not by preconceived criteria. The objectives for use of transformations are those of symmetry, linearity and homoscedasticity. In addition, the use of many resistant techniques such as percentiles and nonparametric test procedures (to be discussed later) are invariant to measurement scale. The results of a rank-sum test, the nonparametric equivalent of a t-test, will be exactly the same whether the original units or logarithms of those units are employed.

1.7.1 The Ladder of Powers

In order to make an asymmetric distribution become more symmetric, the data can be transformed or re-expressed into new units. These new units alter the distances between observations on a line plot. The effect is to either expand or contract the distances to extreme observations on one side of the median, making it look more like the other side. The most commonly-used transformation in water resources is the logarithm. Logs of water discharge, hydraulic conductivity, or concentration are often taken before statistical analyses are performed.

Transformations usually involve power functions of the form $y = x^{\theta}$, where x is the untransformed data, y the transformed data, and θ the power exponent. In figure 1.5 the values of θ are listed in the "ladder of powers" (Velleman and Hoaglin, 1981), a useful structure for determining a proper value of θ.

As can be seen from the ladder of powers, any transformations with θ less than 1 may be used to make right-skewed data more symmetric. Constructing a boxplot or Q-Q plot (see Chapter 2) of the transformed data will indicate whether the transformation was appropriate. Should a logarithmic transformation overcompensate for right skewness and produce a slightly left-skewed distribution, a 'milder' transformation with θ closer to 1, such as a square-root or cube-root transformation, should be employed instead. Transformations with $\theta > 1$ will aid in making left-skewed data more symmetric.

Figure 1.5
"LADDER OF POWERS"
(modified from Velleman and Hoaglin, 1981)

Use	θ	Transformation	Name	Comment
		•		higher powers can be used
		•		
for (−)		•		
skewness	3	x^3	cube	
	2	x^2	square	
	1	x	original units	no transformation
	1/2	$\sqrt{x}$	square root	commonly used
	1/3	$\sqrt[3]{x}$	cube root	commonly used
	0	log(x)	logarithm	commonly used. Holds the place of x^0
for (+) skewness				
	−1/2	$-1/\sqrt{x}$	reciprocal root	the minus sign preserves order of observations
	−1	−1/x	reciprocal	
	−2	$-1/x^2$		
		•		
		•		
		•		lower powers can be used

However, the tendency to search for the 'best' transformation should be avoided. For example, when dealing with several similar data sets, it is probably better to find one transformation which works reasonably well for all, rather than using slightly different ones for each. It must be remembered that each data set is a sample from a larger population, and another sample from the same population will likely indicate a slightly different 'best' transformation. Determination of 'best' in great precision is an approach that is rarely worth the effort.

Exercises

1.1 Yields in wells penetrating rock units without fractures were measured by Wright (1985), and are given below. Calculate the

a) mean
b) trimmed mean
c) geometric mean
d) median
e) compare these estimates of location. Why do they differ?

Unit well yields (in gal/min/ft) in Virginia (Wright, 1985)

0.001	0.030	0.10	0.003	0.040	0.454
0.007	0.041	0.49	0.020	0.077	1.02

1.2 For the well yield data of exercise 1.1, calculate the

a) standard deviation
b) interquartile range
c) MAD
d) skew and quartile skew.

Discuss the differences between a through c.

1.3 Ammonia plus organic nitrogen (in mg/L) was measured in samples of precipitation by Oltmann and Shulters (1989). Some of their data are presented below. Compute summary statistics for these data. Which observation might be considered an outlier? How should this value affect the choice of summary statistics used

a) to compute the mass of nitrogen falling per square mile.
b) to compute a "typical" concentration and variability for these data?

0.3	0.9	0.36	0.92	0.5	1.0
0.7	9.7	0.7	1.3		

Chapter 2

Graphical Data Analysis

Perhaps it seems odd that a chapter on graphics appears at the front of a text on statistical methods. We believe this is very appropriate, as graphs provide crucial information to the data analyst which is difficult to obtain in any other way. For example, figure 2.1 shows eight scatterplots, all of which have exactly the same correlation coefficient. Computing statistical measures without looking at a plot is an invitation to misunderstanding data, as figure 2.1 illustrates. Graphs provide visual summaries of data which more quickly and completely describe essential information than do tables of numbers.

Graphs are essential for two purposes:

1. to provide insight for the analyst into the data under scrutiny, and
2. to illustrate important concepts when presenting the results to others.

The first of these tasks has been called exploratory data analysis (EDA), and is the subject of this chapter. EDA procedures often are (or should be) the 'first look' at data. Patterns and theories of how the system behaves are developed by observing the data through graphs. These are inductive procedures -- the data are summarized rather than tested. Their results provide guidance for the selection of appropriate deductive hypothesis testing procedures.

Once an analysis is complete, the findings must be reported to others. Whether a written report or oral presentation, the analyst must convince the audience that the conclusions reached are supported by the data. No better way exists to do this than through graphics. Many of the same graphical methods which have concisely summarized the information for the analyst will also provide insight into the data for the reader or audience.

The chapter begins with a discussion of graphical methods for analysis of a single data set. Two methods are particularly useful: boxplots and probability plots. Their construction is presented in detail. Next, methods for comparison of two or more groups of data are discussed. Then bivariate plots (scatterplots) are presented, with an especially useful enhancement called a smooth. The chapter ends with a discussion of plots appropriate for multivariate data.

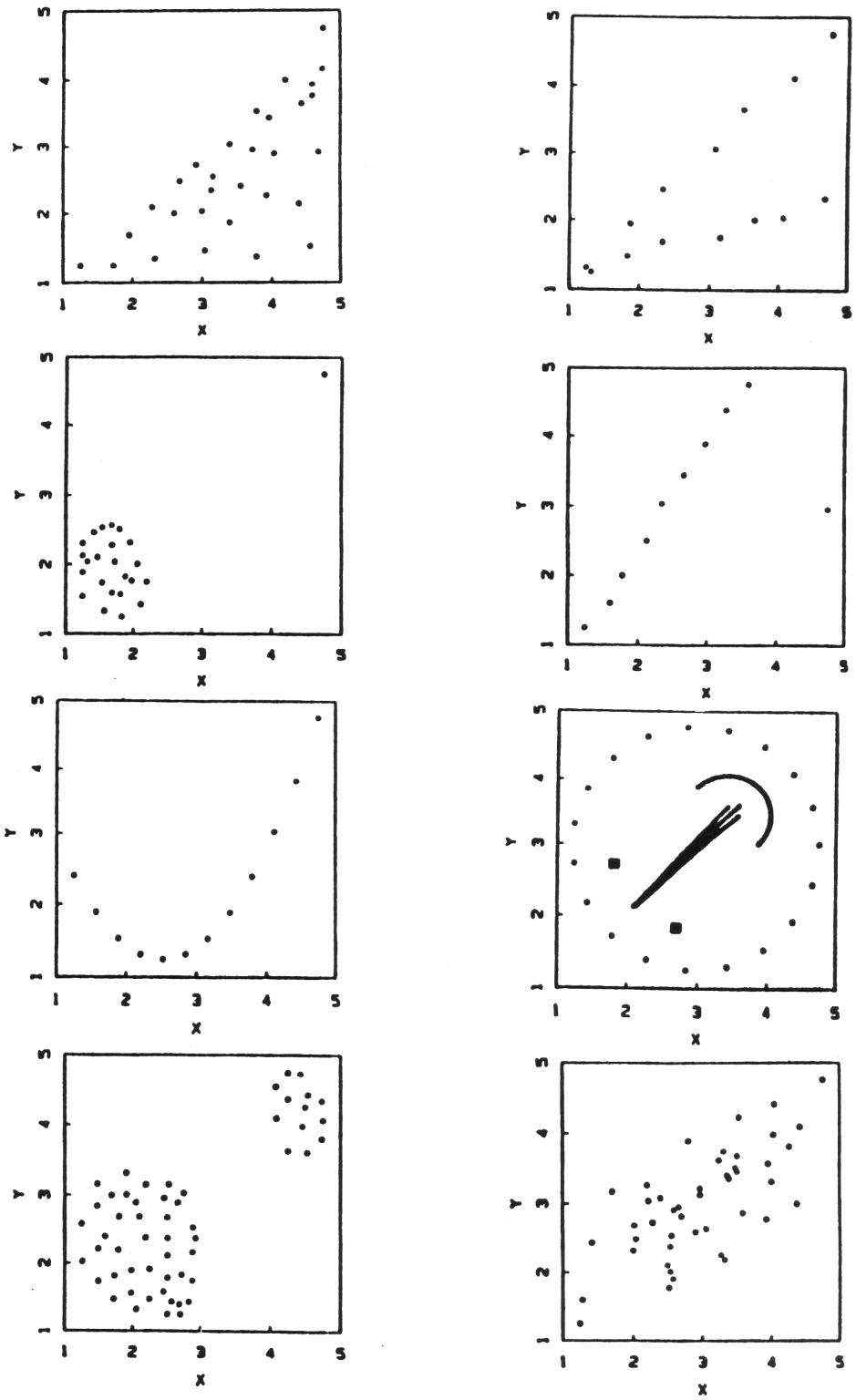


Figure 2.1 Eight scatterplots all with correlation coefficient r = 0.70 (Chambers and others, 1983).

© PWS-Kent Pub. Used with permission.

Throughout sections 2.1 and 2.2 two data sets will be used to compare and contrast the effectiveness of each graphical method. These are annual streamflow (in cubic feet per second, or cfs) for the Licking River at Catawba, Kentucky, from 1929 through 1983, and unit well yields (in gallons per minute per foot of water-bearing material) for valleys without fracturing in Virginia (Wright, 1985).

2.1 Graphical Analysis of Single Data Sets

2.1.1 Histograms

Histograms are familiar graphics, and their construction is detailed in numerous introductory texts on statistics. Bars are drawn whose height is the number n_i, or fraction n_i/n, of data falling into one of several categories or intervals (figure 2.2). Iman and Conover (1983) suggest that, for a sample size of n, the number of intervals k should be the smallest integer such that $2^k \geq n$.

Histograms have one primary deficiency -- their visual impression depends on the number of categories selected for the plot. For example, compare figure 2.2a with 2.2b. Both are histograms of the same data: annual streamflows for the Licking River. Comparisons of shape and similarity among these two figures and the many other possible histograms of the same data depend on the choice of bar widths and centers. False impressions that these are different distributions might be given by characteristics such as the gap around 6,250 cfs. It is seen in 2.2b but not in 2.2a.

Histograms are quite useful for depicting large differences in shape or symmetry, such as whether a data set appears symmetric or skewed. They cannot be used for more precise judgements such as depicting individual values. Thus from figure 2.2a the lowest flow is seen to be larger than 750 cfs, but might be as large as 2,250 cfs. More detail is given in 2.2b, but this lowest observed discharge is still only known to be somewhere between 500 to 1,000 cfs.

For data measured on a continuous scale (such as streamflow or concentration), histograms are not the best method for graphical analysis. The process of forcing continuous data into discrete categories may obscure important characteristics of the distribution. However, histograms are excellent when displaying data which have natural categories or groupings. Examples of such data would include the number of individual organisms found at a stream site grouped by species type, or the number of water-supply wells exceeding some critical yield grouped by geologic unit.

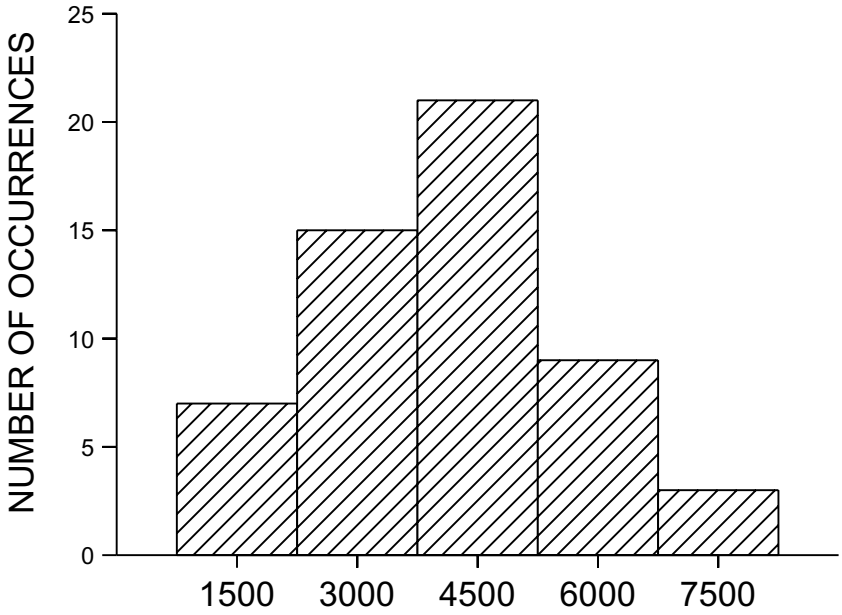


Figure 2.2a. Histogram of annual streamflow for the Licking River

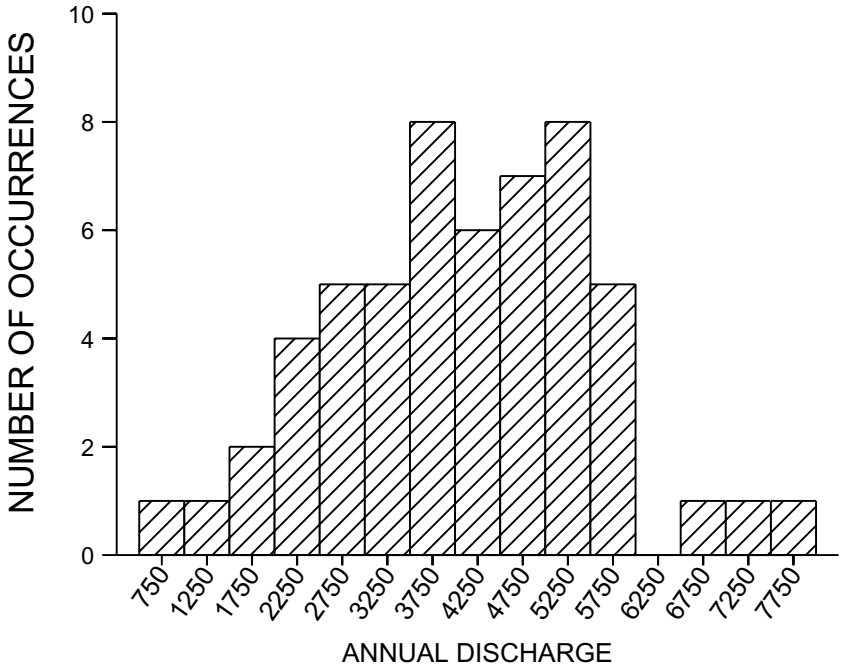


Figure 2.2b. Second histogram of same data, but with different interval divisions.

2.1.2 Stem and Leaf Diagrams

Figure 2.3 shows a stem and leaf (S-L) diagram for the Licking River streamflow data with the same divisions as in figure 2.2b. Stem and leaf diagrams are like histograms turned on their side,

with data magnitudes to two significant digits presented rather than only bar heights. Individual values are easily found. The S-L profile is identical to the histogram and can similarly be used to judge shape and symmetry, but the numerical information adds greater detail. One S-L could function as both a table and a histogram for small data sets.

An S-L is constructed by dividing the range of the data into roughly 10 intervals, and placing the first digit corresponding to these intervals to the left of the vertical line. This is the 'stem', ranging from 0 to 7 (0 to 7000+ cfs) in figure 2.3. Each observation is then represented by one digit to the right of the line (the 'leaves'), so that the number of leaves equals the number of observations falling into that interval. To provide more detail, figure 2.3 has two lines for each stem digit, split to allow 5 leaf digits per line (0-4 and 5-9). Here an asterisk (*) denotes the stem for leaves less than 5, and a period (.) for leaves greater than or equal to 5. For example, in figure 2.3 four observations occur between 2000 and 2500 cfs, with values of 2000, 2200, 2200 and 2400 cfs.

The lowest flow is now seen to be between 900 and l,000 cfs. The gap between 6,000 to 6,500 cfs is still evident, and now the numerical values of the three highest flows are presented. Comparisons between distributions still remain difficult using S-L plots, however, due to the required arbitrary choice of group boundaries.

```
(range in cfs)
( 500- 999)          +0. | 9
(1000-1499)           1* | 2
(1500-1999)           1. | 59
(2000-2499)           2* | 0224
(2500-2999)           2. | 66889
(3000-3499)           3* | 01122
(3500-3999)           3. | 55678889
(4000-4499)           4* | 000124
(4500-4999)           4. | 5566777
(5000-5499)           5* | 01123334
(5500-5999)           5. | 56899
(6000-6499)           6* |
(6500-6999)           6. | 8
(7000-7499)           7* | 2
(7500-7999)           7. | 7
```

Figure 2.3 Stem and Leaf Plot of Annual Streamflow
(Leaf digit unit = 100 1|2 represents 1200)

2.1.3 Quantile Plots

Quantile plots visually portray the quantiles, or percentiles (which equal the quantiles times 100) of the distribution of sample data. Quantiles of importance such as the median are easily discerned (quantile, or cumulative frequency = 0.5). With experience, the spread and skewness of the data, as well as any bimodal character, can be examined. Quantile plots have three advantages:

1. Arbitrary categories are not required, as with histograms or S-L's.
2. All of the data are displayed, unlike a boxplot.
3. Every point has a distinct position, without overlap.

Figure 2.4 is a quantile plot of the streamflow data from figure 2.2. Attributes of the data such as the gap between 6000 and 6800 cfs (indicated by the nearly horizontal line segment) are evident. The percent of data in the sample less than a given cfs value can be read from the graph with much greater accuracy than from a histogram.

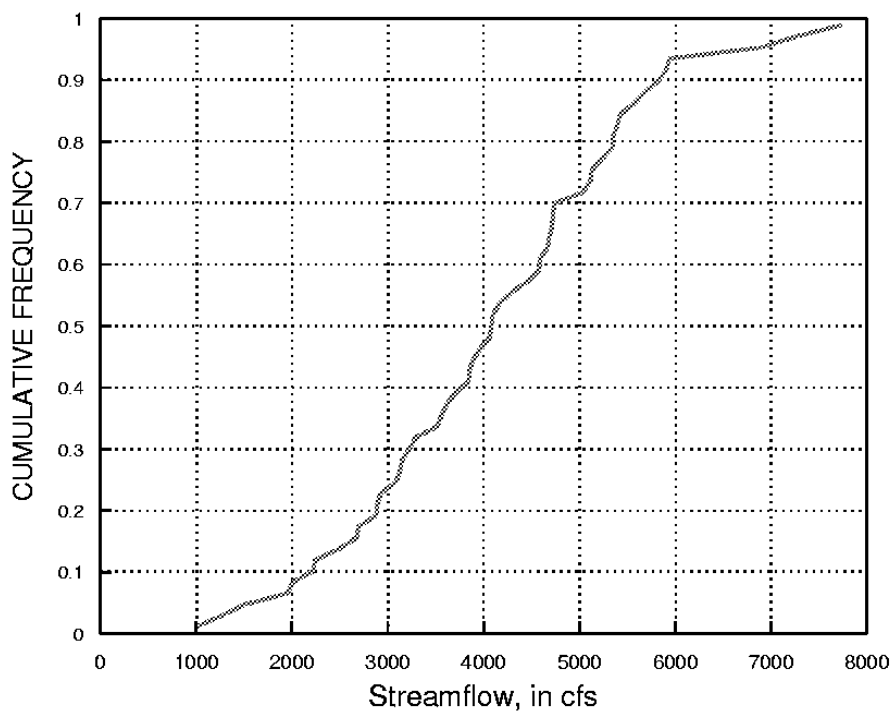


Figure 2.4 Quantile plot of the Licking R. annual streamflow data

2.1.3.1 Construction of a quantile plot

To construct a quantile plot, the data are ranked from smallest to largest. The smallest data value is assigned a rank i=1, while the largest receives a rank i=n, where n is the sample size of the data set. The data values themselves are plotted along one axis, usually the horizontal axis. On the other axis is the "plotting position", which is a function of the rank i and sample size n. As discussed in the next section, the Cunnane plotting position $p_i = (i–0.4)/(n+0.2)$ is used in

this book. Below are listed the first and last 5 of the 55 data pairs used in construction of figure 2.4. When tied data values are present, each is assigned a separate plotting position (the plotting positions are not averaged). In this way tied values are portrayed as a vertical "cliff" on the plot.

q_i = Licking R. streamflow, in cfs p_i = plotting position

i	q_i	p_i	i	q_i	p_i	i	q_i	p_i
1	994.3	.01	5	2006.0	.08	52	5937.3	.93
2	1263.1	.03		•		53	6896.0	.95
3	1504.2	.05		•		54	7270.1	.97
4	1949.5	.07	51	5907.0	.92	55	7730.7	.99

Quantile plots are sample approximations of the cumulative distribution function (cdf) of a continuous random variable. The cdf for a normal distribution is shown in figure 2.7. A second approximation is the sample (or empirical) cdf, which differs from quantile plots in its vertical scale. The vertical axis of a sample cdf is the probability i/n of being less than or equal to that observation. The largest observation has $i/n = 1$, and so has a zero probability of being exceeded. For samples (subsets) taken from a population, a nonzero probability of exceeding the largest value observed thus far should be recognized. This is done by using the plotting position, a value less than i/n, on the vertical axis of the quantile plot. As sample sizes increase, the quantile plot will more closely mimic the underlying population cdf.

2.1.3.2 Plotting positions

Variations of quantile plots are used frequently for three purposes:

1. to compare two or more data distributions (a Q-Q plot),
2. to compare data to a normal distribution (a probability plot), and
3. to calculate frequencies of exceedance (a flow-duration curve).

Unfortunately, different plotting positions have traditionally been used for each of the above three purposes. It would be desirable instead to use one formula that is suitable for all three. Numerous plotting position formulas have been suggested, most having the general formula

$$p = (i - a) / (n + 1 - 2a)$$

where a varies from 0 to 0.5. Five of the most commonly-used formulas are:

Reference	a	Formula
Weibull (1939)	0	$i / (n + 1)$
Blom (1958)	0.375	$(i - 0.375) / (n + 0.25)$
Cunnane (1978)	0.4	$(i - 0.4) / (n + 0.2)$
Gringorten (1963)	0.44	$(i - 0.44) / (n + 0.12)$
Hazen (1914)	0.5	$(i - 0.5) / n$

The Weibull formula has long been used by hydrologists in the United States for plotting flow-duration and flood-frequency curves (Langbein, 1960). It is used in Bulletin 17B, the standard

reference for determining flood frequencies in the United States (Interagency Advisory Committee on Water Data, 1982). The Blom formula is best for comparing data quantiles to those of a normal distribution in probability plots, though all of the above formulas <u>except</u> the Weibull are acceptable for that purpose (Looney and Gulledge, 1985b). The Hazen formula is used by Chambers and others (1983) for comparing two or more data sets using Q-Q plots.

Separate formulae could be used for the situations in which each is optimal. In this book we instead use one formula, the Cunnane formula given above, for all three purposes. We do this in an attempt to simplify. The Cunnane formula was chosen because

1. it is acceptable for normal probability plots, being very close to Blom.
2. it is used by Canadian and some European hydrologists for plotting flow-duration and flood-frequency curves. Cunnane (1978) presents the arguments for use of this formula over the Weibull when calculating exceedance probabilities.

For convenience when dealing with small sample sizes, table B1 of the Appendix presents Cunnane plotting positions for sample sizes n = 5 to 20.

2.1.4 Boxplots

A very useful and concise graphical display for summarizing the distribution of a data set is the boxplot (figure 2.5). Boxplots provide visual summaries of

1) the center of the data (the median--the center line of the box)
2) the variation or spread (interquartile range--the box height)
3) the skewness (quartile skew--the relative size of box halves)
4) presence or absence of unusual values ("outside" and "far outside" values).

Boxplots are even more useful in comparing these attributes among several data sets.

Compare figures 2.4 and 2.5, both of the Licking R. data. Boxplots do not present all of the data, as do stem-and-leaf or quantile plots. Yet presenting all data may be more detail than is necessary, or even desirable. Boxplots do provide concise visual summaries of essential data characteristics. For example, the symmetry of the Licking R. data is shown in figure 2.5 by the similar sizes of top and bottom box halves, and by the similar lengths of whiskers. In contrast, in figure 2.6 the taller top box halves and whiskers indicate a right-skewed distribution, the most commonly occurring shape for water resources data. Boxplots are often put side-by-side to visually compare and contrast groups of data.

Three commonly used versions of the boxplot are described as follows (figure 2.6 a,b, and c). Any of the three may appropriately be called a boxplot.

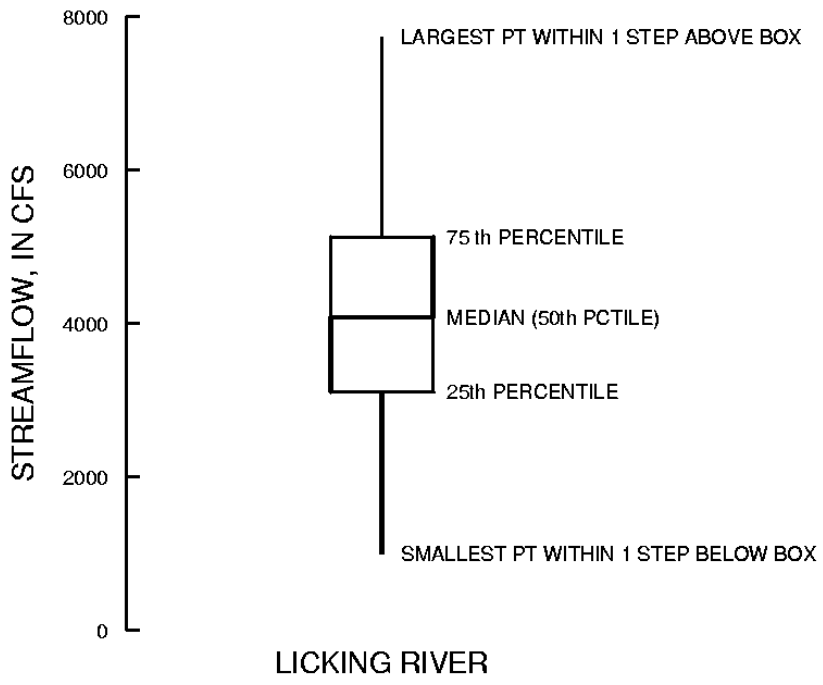


Figure 2.5 Boxplot for the Licking R. data

2.1.4.1 Simple boxplot

The simple boxplot was originally called a "box-and-whisker" plot by Tukey (1977). It consists of a center line (the median) splitting a rectangle defined by the upper and lower hinges (very similar to quartiles -- see appendix). Whiskers are lines drawn from the ends of the box to the maximum and minimum of the data, as depicted in graph a of figure 2.6.

2.1.4.2 Standard boxplot

Tukey's "schematic plot" has become the most commonly used version of a boxplot (graph b in figure 2.6), and will be the type of boxplot used throughout this book. With this standard boxplot, outlying values are distinguished from the rest of the plot. The box is as defined above. However, the whiskers are shortened to extend only to the last observation within one step beyond either end of the box ("adjacent values"). A step equals 1.5 times the height of the box (1.5 times the interquartile range). Observations between one and two steps from the box in either direction, if present, are plotted individually with an asterisk ("outside values"). Outside values occur fewer than once in 100 times for data from a normal distribution. Observations farther than two steps beyond the box, if present, are distinguished by plotting them with a small circle ("far-out values"). These occur fewer than once in 300,000 times for a normal distribution. The occurrence of outside or far-out values more frequently than expected gives a quick visual indication that data may not originate from a normal distribution.

2.1.4.3 Truncated boxplot

In a third version of the boxplot (graph c of figure 2.6), the whiskers are drawn only to the 90th and 10th percentiles of the data set. The largest 10 percent and smallest 10 percent of the data are not shown. This version could easily be confused with the simple boxplot, as no data appear beyond the whiskers, and should be clearly defined as having eliminated the most extreme 20 percent of data. It should be used only when the extreme 20 percent of data are not of interest.

In a variation on the truncated boxplot, Cleveland (1985) plotted all observations beyond the 10th and 90th percentile-whiskers individually, calling this a "box graph". The weakness of this style of graph is that 10 percent of the data will always be plotted individually at each end, and so the plot is far less effective than a standard boxplot for defining and emphasizing unusual values.

Further detail on construction of boxplots may be found in the appendix, and in Chambers and others (1983) and McGill and others (1978).

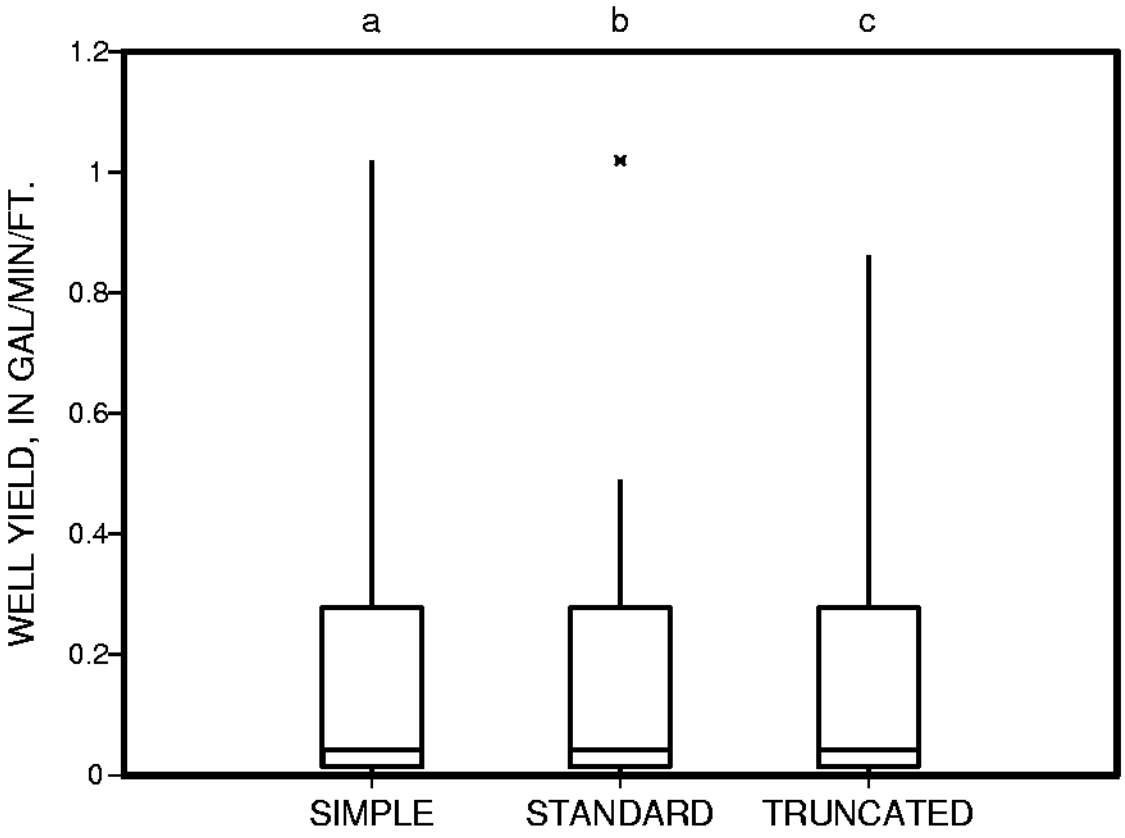


Figure 2.6 Three versions of the boxplot (unit well yield data).

2.1.5 Probability Plots

Probability plots are used to determine how well data fit a theoretical distribution, such as the normal, lognormal, or gamma distributions. This could be attempted by visually comparing

histograms of sample data to density curves of the theoretical distributions such as figures 1.1 and 1.2. However, research into human perception has shown that departures from straight lines are discerned more easily than departures from curvilinear patterns. By expressing the theoretical distribution as a straight line, departures from the distribution are more easily perceived. This is what occurs with a probability plot.

To construct a probability plot, quantiles of sample data are plotted against quantiles of the standardized theoretical distribution. In figure 2.7, quantiles from the quantile plot of the Licking R. streamflow data (lower scale) are overlain with the S-shaped quantiles of the standard normal distribution (upper scale). For a given cumulative frequency (plotting position, p), quantiles from each curve are paired and plotted as one point on the probability plot, figure 2.8. Note that quantiles of the data are simply the observation values themselves, the pth quantiles where $p = (i–0.4)/(n+0.2)$. Quantiles of the standard normal distribution are available in table form in most textbooks on statistics. Thus, for each observation, a pair of quantiles is plotted in figure 2.8 as one point. For example, the median (p=0.5) equals 0 for the standard normal, and 4079 cfs for the Licking R. data. The point (0,4079) is one point included in figure 2.8. Data closely approximating the shape of the theoretical distribution, in this case a normal distribution, will plot near to a straight line.

To illustrate the construction of a probability plot in detail, data on unit well yields (y_i) from Wright (1985) will be plotted versus their normal quantiles (also called normal scores). The data are ranked from the smallest (i=1) to largest (i=n), and their corresponding plotting positions $p_i = (i – 0.4)/(n + 0.2)$ calculated. Normal quantiles (Zp) for a given plotting position p_i may be obtained in one of three ways:

a. from a table of the standard normal distribution found in most statistics textbooks
b. from table B2 in the Appendix, which presents standard normal quantiles for the Cunnane plotting positions of table B1
c. from a computerized approximation to the inverse standard normal distribution available in many statistical packages, or as listed by Zelen and Severo (1964).

Entering the table with $p_i = .05$, for example, will provide a $Zp = –1.65$. Note that since the median of the standard normal distribution is 0, Zp will be symmetrical about the median, and only half of the Zp values must be looked up:

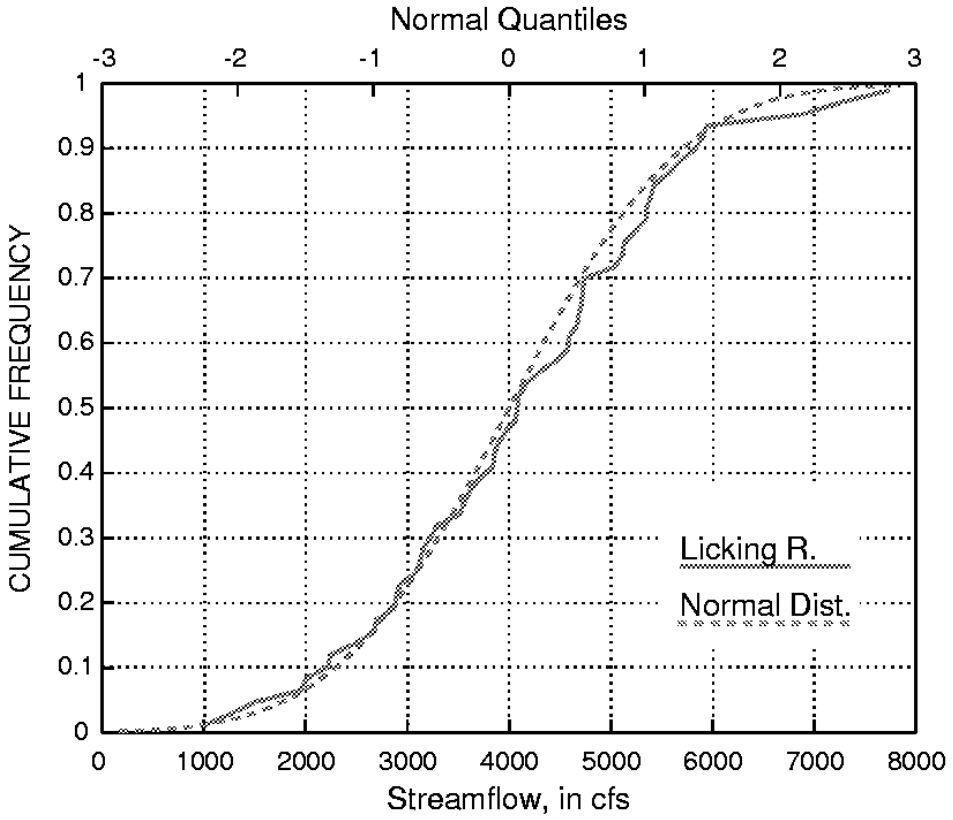


Figure 2.7 Overlay of Licking R. and standard normal distribution quantile plots

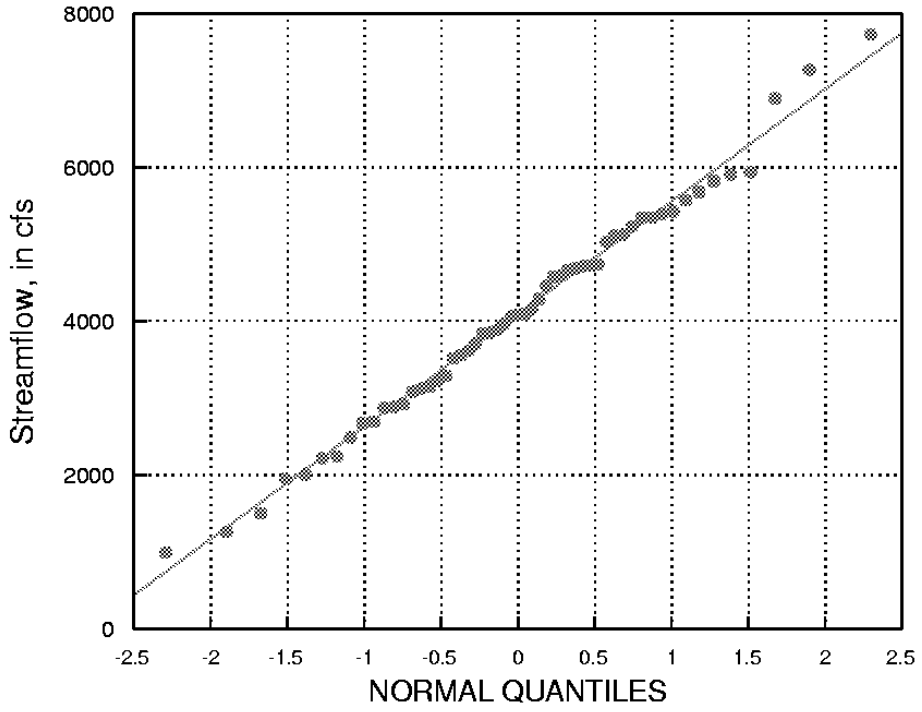


Figure 2.8 Probability plot of the Licking R. data

Unit well yields (in gal/min/ft) for valleys without fracturing (Wright, 1985)
y_i = yield p_i = plotting position Zp = normal quantile of p

i	y_i	p_i	Zp	i	y_i	p_i	Zp	i	y_i	p_i	Zp
1	0.001	.05	–1.65	5	0.030	.38	–.31	9	0.10	.70	.52
2	0.003	.13	–1.13	6	0.040	.46	–.10	10	0.454	.79	.80
3	0.007	.21	–0.80	7	0.041	.54	.10	11	0.49	.87	1.13
4	0.020	.30	–0.52	8	0.077	.62	.31	12	1.02	.95	1.65

For comparison purposes, it is helpful to plot a reference straight line on the plot. The solid line on figure 2.8 is the normal distribution which has the same mean and standard deviation as do the sample data. This reference line is constructed by plotting $\bar{y}$ as the y intercept of the line (Zp=0), so that the line is centered at the point $(0, \bar{y})$, the mean of both sets of quantiles. The standard deviation s is the slope of the line on a normal probability plot, as the quantiles of a standard normal distribution are in units of standard deviation. Thus the line connects the points $(0, \bar{y})$ and $(1 , \bar{y} + s)$.

2.1.5.1 Probability paper

Specialized 'probability paper' is often used for probability plots. This paper simply retransforms the linear scale for quantiles of the standard distribution back into a nonlinear scale of plotting positions (figure 2.9). There is no difference between the two versions except for the horizontal scale. With probability paper the horizontal axis can be directly interpreted as the percent probability of occurrence, the plotting position times 100. The linear quantile scale of figure 2.8 is sometimes included on probability paper as 'probits,' where a probit = normal quantile + 5.0. Probability paper is available for distributions other than the normal, but all are constructed the same way, using standardized quantiles of the theoretical distribution.

In figure 2.9 the lower horizontal scale results from sorting the data in increasing order, and assigning rank 1 to the smallest value. This is commonly done in water-quality and low-flow studies. Had the data been sorted in decreasing order, assigning rank 1 to the largest value as is done in flood-flow studies, the upper scale would result -- the percent exceedance. Either horizontal scale may be obtained by subtracting the other from 100 percent.

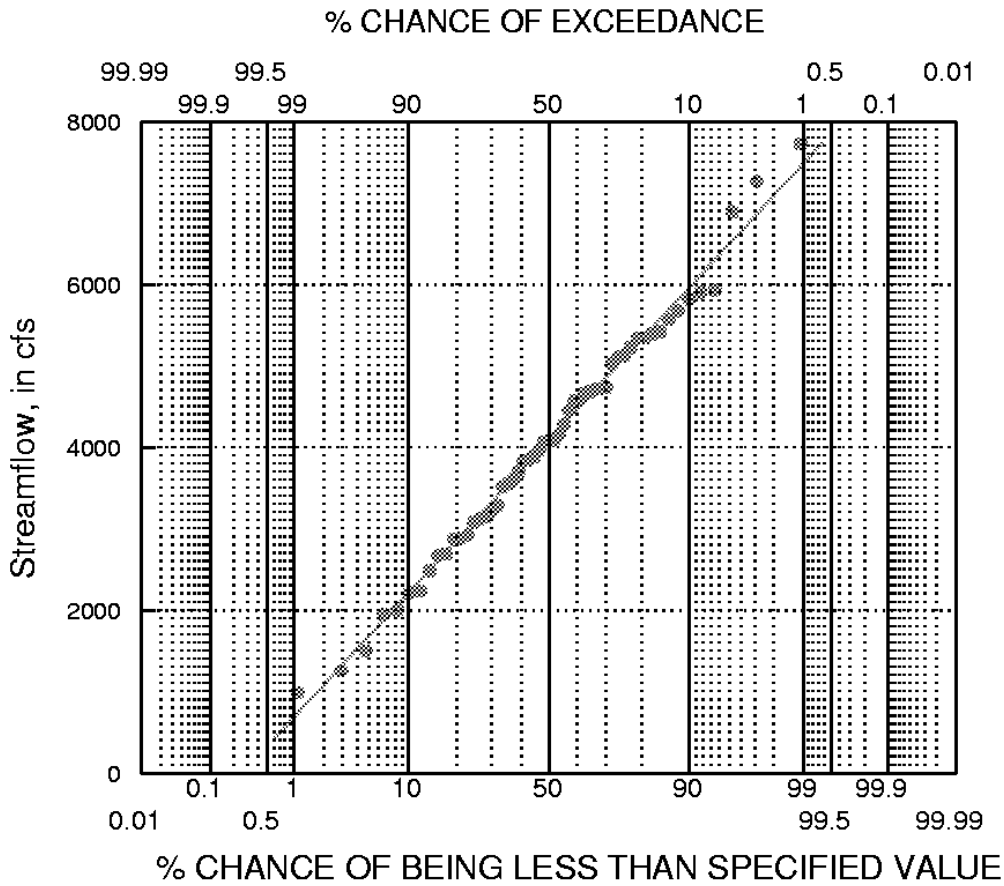


Figure 2.9 -- Probability plot of Licking R. data on probability paper

2.1.5.2 Deviations from a linear pattern

If probability plots do not exhibit a linear pattern, their nonlinearity will indicate why the data do not fit the theoretical distribution. This is additional information that hypothesis tests for normality (described later) do not provide. Three typical conditions resulting in deviations from linearity are: asymmetry or skewness, outliers, and heavy tails of the distribution. These are discussed below.

Figure 2.10 is a probability plot of the base 10 logarithms of the Licking R. data. The data are negatively (left) skewed. This is seen in figure 2.10 as a greater slope on the left-hand side of the plot, producing a slightly convex shape. Figure 2.11 shows a right-skewed distribution, the unit well yield data. The lower bound of zero, and the large slope on the right-hand side of the plot produces an overall concave shape. Thus probability plots can be used to indicate what type of transformation is needed to produce a more symmetric distribution. The degree of curvature gives some indication of the severity of skewness, and therefore the degree of transformation required.

Outliers appear on probability plots as departures from the pattern of the rest of the data. Figure 2.12 is a probability plot of the Licking R. data, but the two largest observations have been altered (multiplied by 3). Compare figures 2.12 and 2.8. Note that the majority of points in figure 2.12 still retain a linear pattern, with the two outliers offset from that pattern. Note that the straight line, a normal distribution with mean and standard deviation equal to those of the altered data, does not fit the data well. This is because the mean and standard deviation are inflated by the two outliers.

The third departure from linearity occurs when more data are present in both tails (areas furthest from the median) than would be expected for a normal distribution. Figure 2.13 is a probability plot of adjusted nitrate concentrations in precipitation from Wellston, Michigan (Schertz and Hirsch, 1985). These data are actually residuals (departures) from a regression of log of nitrate concentration versus log of precipitation volume. A residual of 0 indicates that the concentration is exactly what would be expected for that volume, a positive residual more than what is expected, and negative less than expected. The data in figure 2.13 display a predominantly linear pattern, yet one not fit well by the theoretical normal shown as the solid line. Again this lack of fit indicates outliers are present. The outliers are data to the left which plot below the linear pattern, and those above the pattern to the right of the figure. Outliers occur on both ends in greater numbers than expected from a normal distribution. A boxplot for the data is shown in figure 2.14 for comparison. Note that both the box and whiskers are symmetric, and therefore no power transformation such as those in the "ladder of powers" would produce a more nearly normal distribution. Data may depart from a normal distribution not only in skewness, but by the number of extreme values. Excessive numbers of extreme values may cause significance levels of tests requiring the normality assumption to be in error. Therefore procedures which assume normality for their validity when applied to data of this type may produce quite inaccurate results.

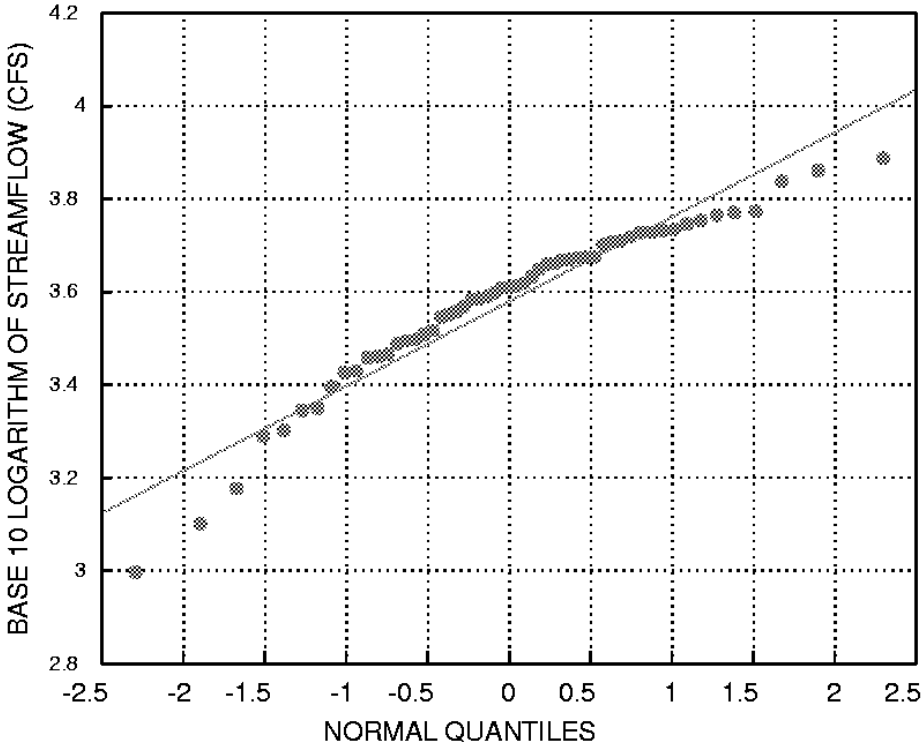


Figure 2.10 -- Probability plot of a left-skewed distribution (logs of Licking R. data)

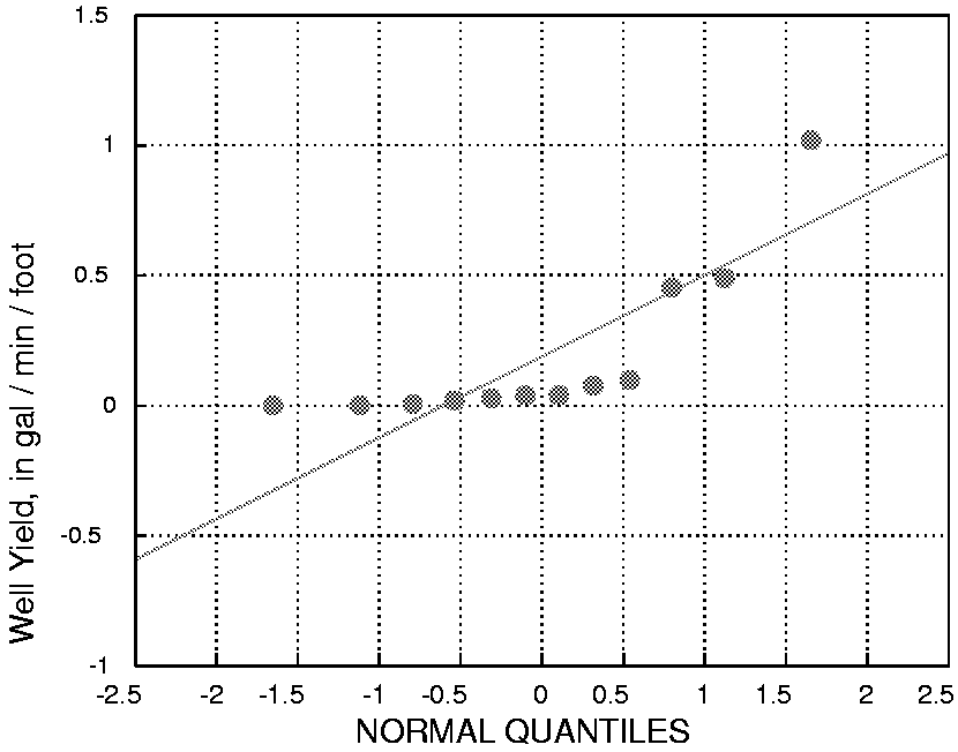


Figure 2.11 -- Probability plot of a right-skewed distribution (unit well yields)

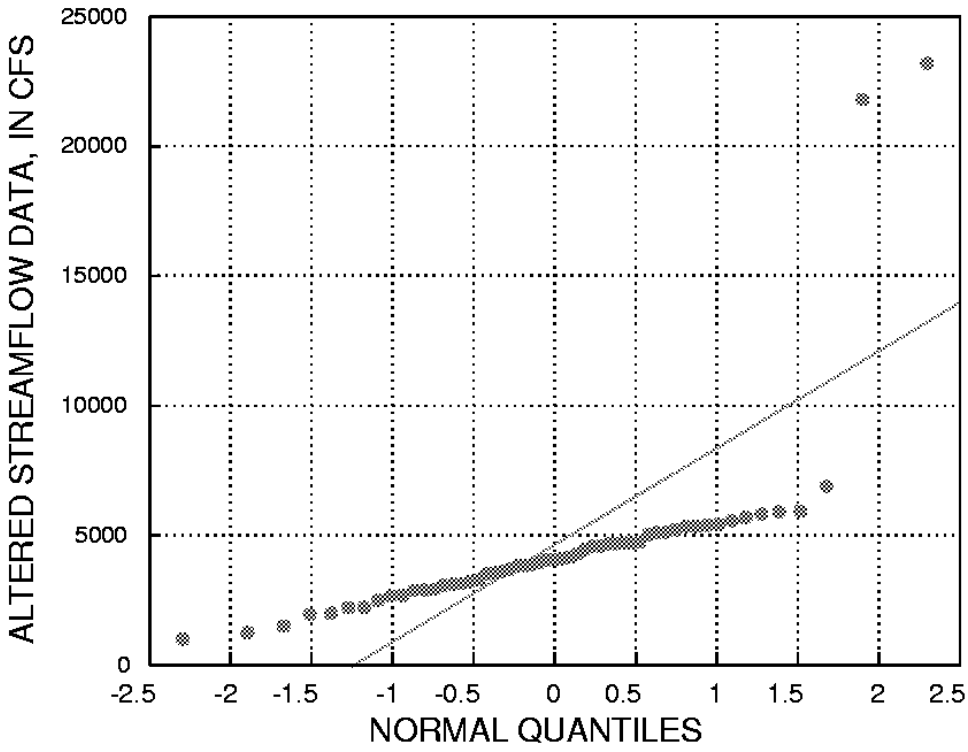


Figure 2.12 -- Probability plot of data with high outliers

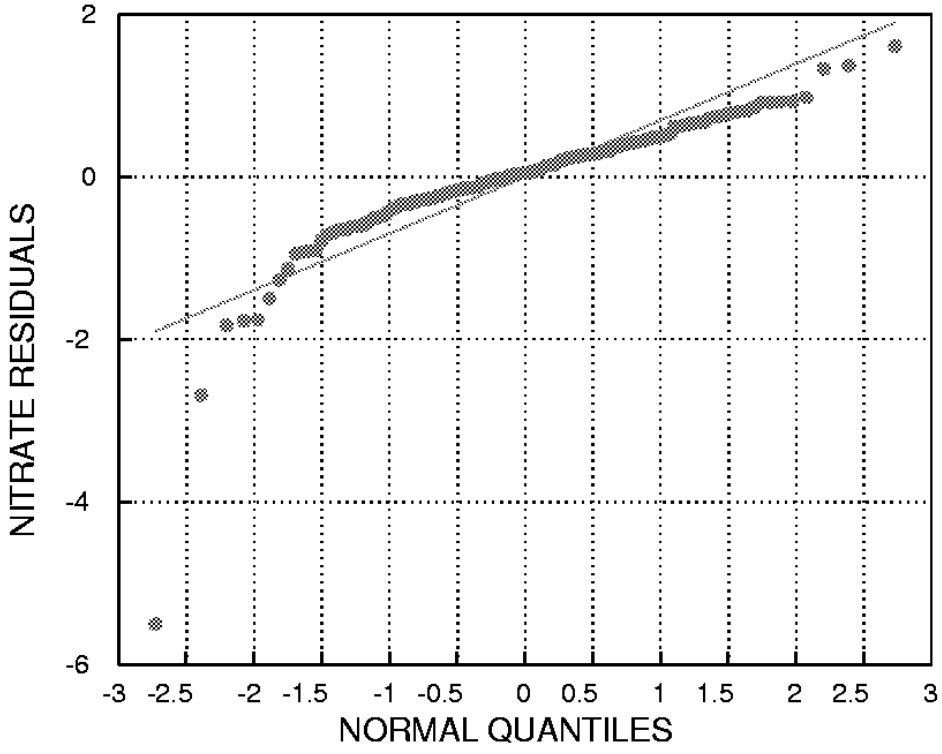


Figure 2.13 -- Probability plot of a heavy-tailed data set

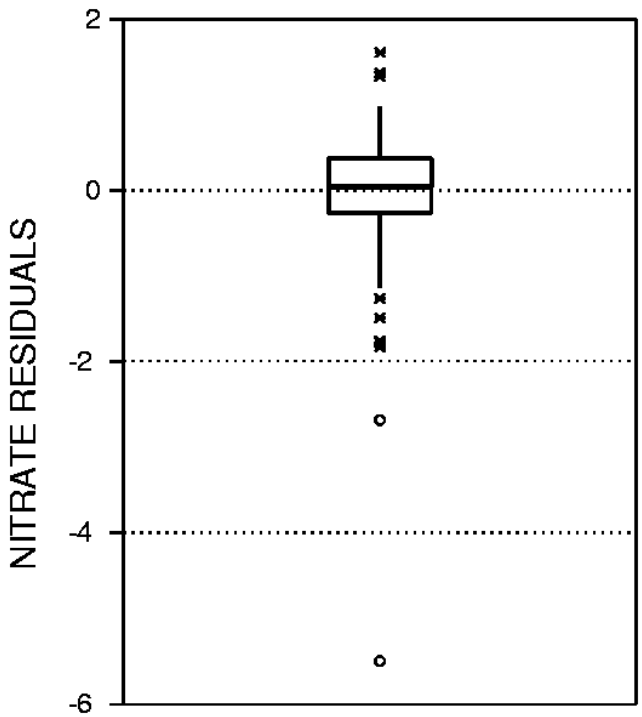


Figure 2.14 -- Boxplot of a heavy-tailed data set

2.1.5.3 Probability plots for comparing among distributions

In addition to comparisons to a normal distribution, quantiles may be computed and probability plots constructed for any two-parameter distribution. The distribution which causes data to be most like a straight line on its probability plot is the one which most closely resembles the distributional shape of the data. Data may be compared to a two-parameter lognormal distribution by simply plotting the logarithms of the data as the data quantiles, as was done in figure 2.10. Vogel (1986) demonstrated the construction of probability plots for the Gumbel (extreme-value) distribution, which is sometimes employed for flood-flow studies. Vogel and Kroll (1989) cover the use of probability plots for the two-parameter Weibull distribution, used in fitting low-flow data. Again, the best fit is obtained with the distribution which most closely produces a linear plot. In both references, the use of a test of significance called the probability plot correlation coefficient augmented the visual determination of linearity on the plot. This test will be covered in detail in Chapter 4.

Use of three-parameter distributions can also be indicated by probability plots. For example, if significant right-skewness remains after logarithms are taken, the resulting concave shape on a lognormal probability plot indicates that a log-Pearson III distribution would better fit the data. Vogel and Kroll (1989) demonstrate the construction of a probability plot for the log-Pearson III distribution using a Wilson-Hilferty transformation.

2.2 Graphical Comparisons of Two or More Data Sets

Each of the graphical methods discussed thus far can be, and have been, used for comparing more than one group of data. However, each is not equally effective. As the following sections show, histograms are not capable of providing visual comparisons between data sets at the same level of detail as boxplots or probability plots. Boxplots excel in clarity and easy discrimination of important distributional characteristics, even for comparisons between many groups of data. A newer type of plot, the quantile-quantile (Q-Q) plot, provides additional information about the relationship between two data sets.

Each graphic will be developed for the same data set, a comparison of unit well yields in Virginia (Wright, 1985). These are small data sets: 13 wells are from valleys underlain by fractured rocks, and 12 wells from valleys underlain by unfractured rocks.

2.2.1 Histograms

Figure 2.15 presents histograms for the two sets of well yield data. The right-skewness of each data set is easily seen, but it is difficult to discern whether any differences exist between them. Histograms do not provide a good visual picture of the centers of the distributions, and only a slightly better comparison of spreads. Positioning histograms side-by-side instead of one above the other provide even less ability to compare data, as the data axes would not be aligned. Unfortunately, this is commonly done. Also common are overlapping histograms, such as in figure 2.16. Overlapping histograms provide poor visual discrimination between multiple data sets.

2.2.2 Dot and Line Plots of Means, Standard Deviations

Figure 2.17 is a "dot and line" plot often used to represent the mean and standard deviation (or standard error) of data sets. Each dot is the mean of the data set. The bars extend to plus and minus either one standard deviation (shown), or plus and minus one or more standard errors (s.e. = $s/\sqrt{n}$), beyond the mean. This plot displays differences in mean yields, but little else. No information on the symmetry of the data or presence of outliers is available. Because of this, there is not much information given on the spread of the data, as the standard deviation may describe the spread of most of the data, or may be strongly influenced by skewness and a few outliers.

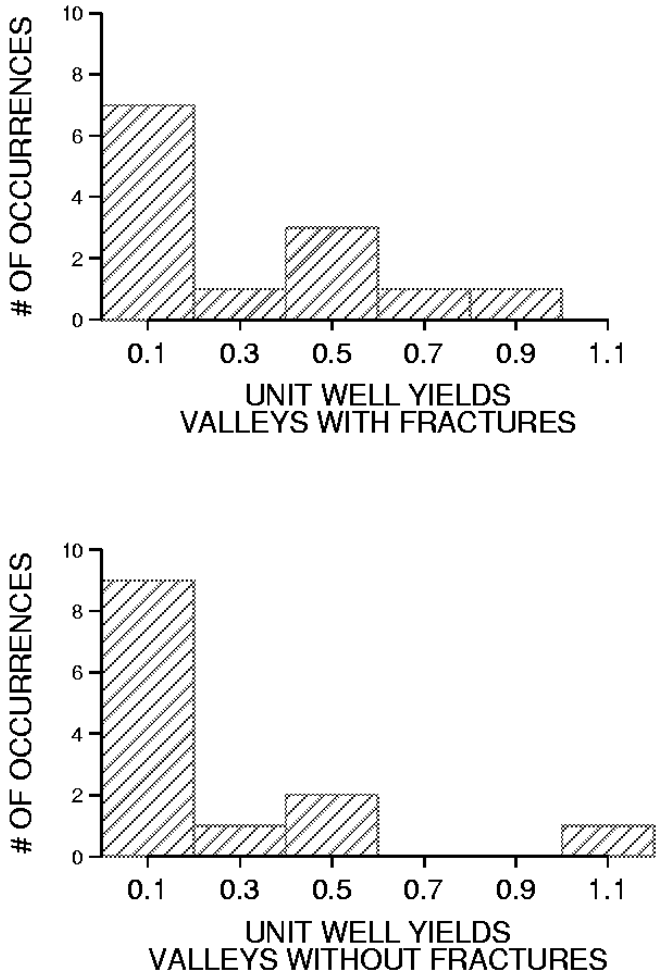


Figure 2.15 Histograms of the unit well yield data

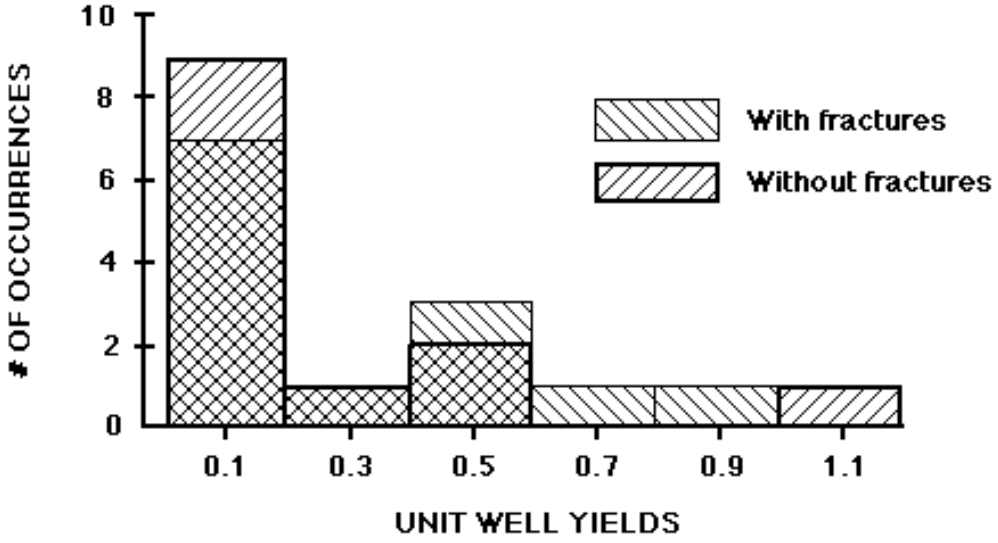


Figure 2.16 Overlapping histograms of the unit well yield data

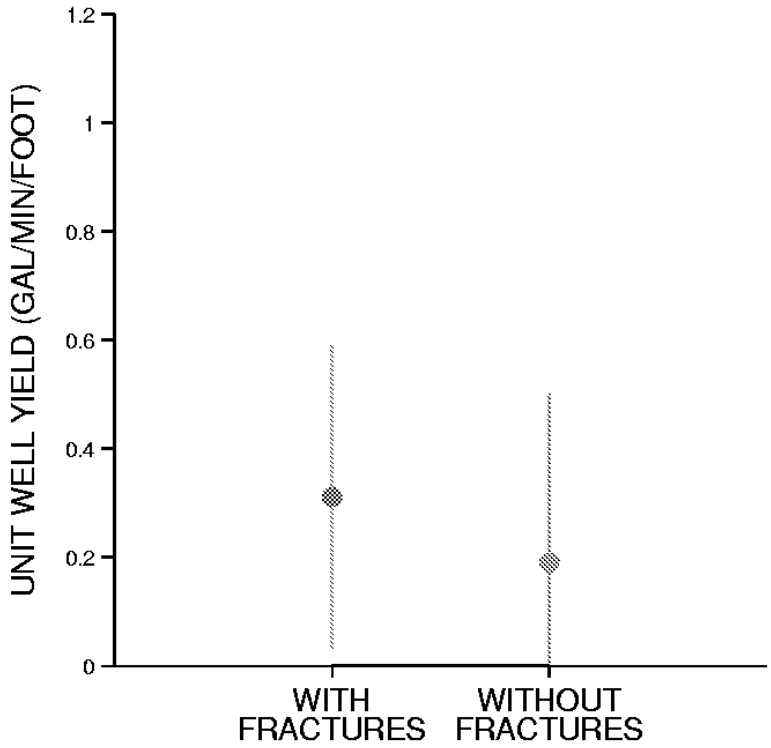


Figure 2.17 Dot and line plot for the unit well yield data

To emphasize the deficiencies of dot and line plots such as these, figure 2.18 presents three data sets with very different characteristics. The first is a uniform distribution of values between 0 and 20. It is symmetric. The second is a right-skewed data set with outliers. The third is a bimodal distribution, also symmetric. All three have a mean of 10 and standard deviation of 6.63. Therefore each of the three would be represented by the same dot and line plot, shown at the right of the figure.

Dot and line plots are useful only when the data are actually symmetric. If skewness or outliers are present, as with data set 2, neither the plots (or a table of means and standard deviations) indicate their presence. Even for symmetric distributions, differences such as those between data sets 1 and 3 will not be evident. Far better graphical methods are available.

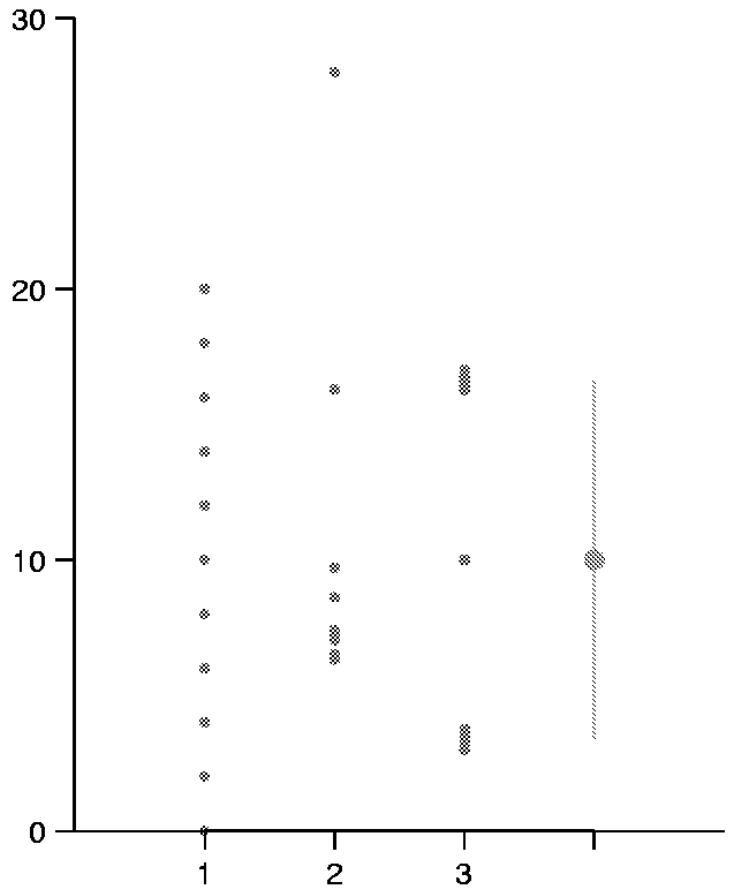


Figure 2.18 Number lines of 3 dissimilar groups of data, all having an identical dot and line plot (shown at right).

2.2.3 Boxplots

Figure 2.19 presents boxplots of the well yield data. The median well yield is seen to be higher for the areas with fractures. The IQR of wells with fractures is slightly larger than that for wells without, and the highest value for each group is similar. Both data sets are seen to be right-skewed. Thus a large amount of information is contained in this very concise illustration. The mean yield, particularly for wells without fractures, is undoubtedly inflated due to skewness, and differences between the two groups of data will in general be larger than indicated by the differences in their mean values.

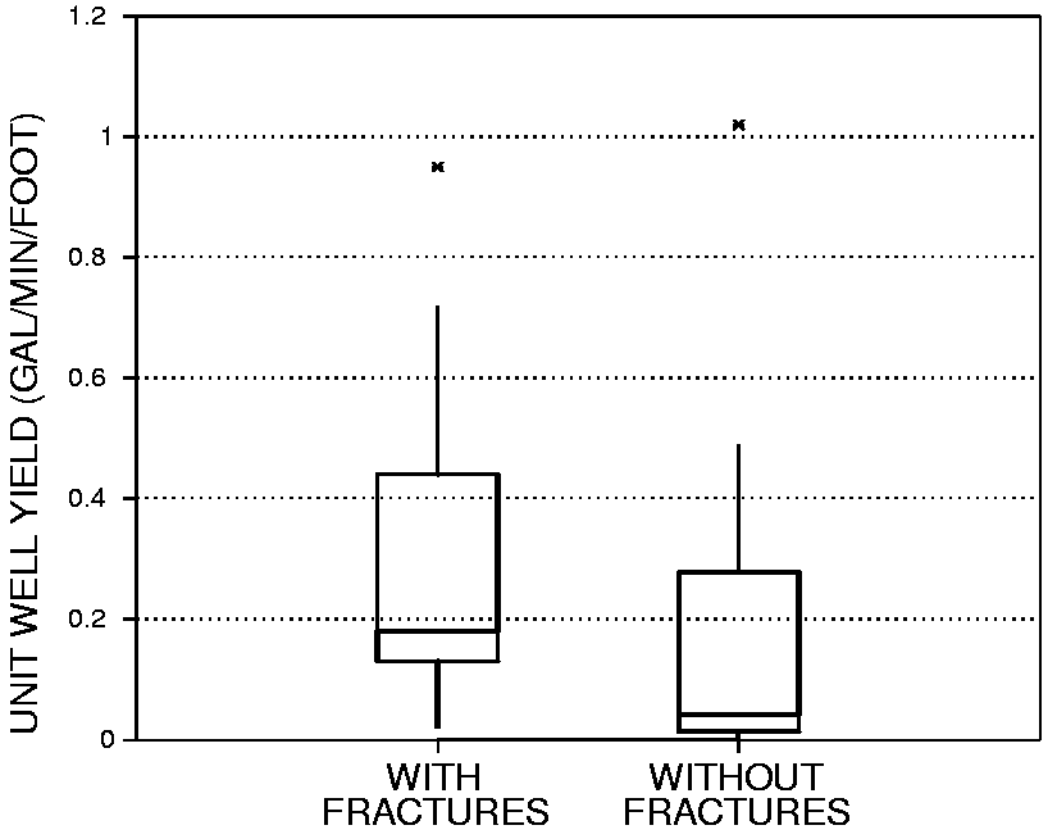


Figure 2.19 Boxplots of the unit well yield date

In figure 2.20, boxplots of the three data sets given in figure 2.18 are presented. The skewness of data set 2 is clear, as is the symmetry of 1 and 3. The difference in shape between 1 and 3 is evident. The minute whiskers of data set 3 illustrate that over 25 percent of the data are located essentially at the upper and lower quartiles -- a bimodal distribution.

The characteristics which make boxplots useful for inspecting a single data set make them even more useful for comparing multiple data sets. They are valuable guides in determining whether central values, spread, and symmetry differ among groups of data. They will be used in later chapters to guide whether tests based on assumptions of normality may be employed. The essential characteristics of numerous groups of data may be displayed in a small space. For example, the 20 boxplots of figure 2.21 were used by Holtschlag (1987) to illustrate the source of ammonia nitrogen on a section of the Detroit River. The Windmill Point Transect is upstream of the U. S. city of Detroit, while the Fermi Transect is below the city. Note the marked changes in concentration (the median lines of the boxplots) and variability (the widths of the boxes) on the Michigan side of the river downstream of Detroit. A lot of information on streamwater quality is succinctly summarized in this relatively small figure.

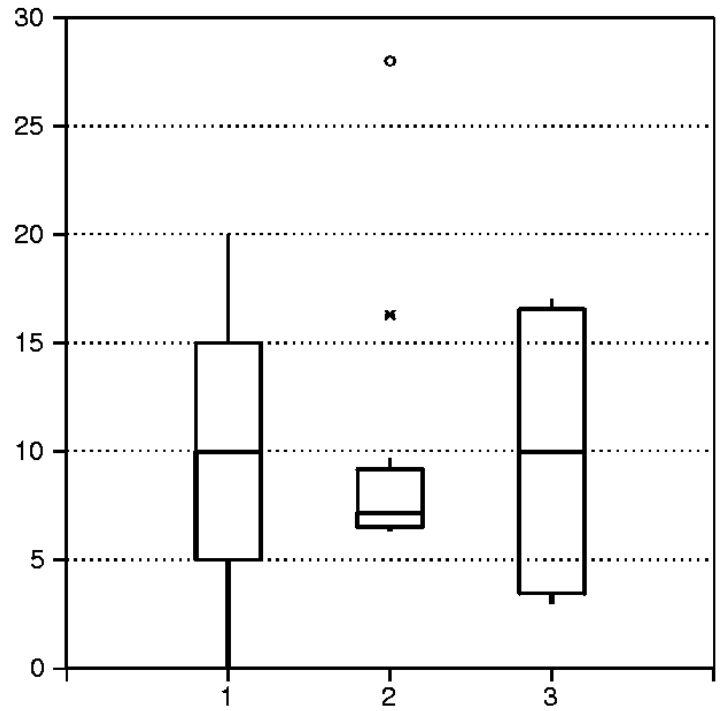


Figure 2.20 Boxplots of the 3 dissimilar groups of data shown in figure 2.18

2.2.4 Probability Plots

Probability plots are also useful graphics for comparing groups of data. Characteristics evident in boxplots are also seen using probability plots, though in a different format. Comparisons of each quantile, not just the boxplot quartiles, can be made. The straightness of each data set also allows quick comparisons to conformity with the theoretical distribution.

Figure 2.22 is a probability plot of the two well yield data sets. The right-skewness of each data set is shown by their concave shapes. Wells without fractures have greater skewness as shown by their greater concavity on the plot. Quantiles of the wells with fractures are higher than those without, indicating generally higher yields. Figure 2.22 shows that the lowest yields in each group are similar, as both data sets approach zero yield. Also seen are the similarity in the highest yield for each group, due to the outlier for the without fractures group. Comparisons between median values are simple to do -- just travel up the normal quantile = 0 line. Comparisons of spreads are more difficult -- the slopes of each data set display their spread.

In general, boxplots summarize the differences between data groups in a manner more quickly discerned by the viewer. When comparisons to a particular theoretical distribution such as the normal are important, or comparisons between quantiles other than the quartiles are necessary, probability plots are useful graphics. Either have many advantages over histograms or dot and line plots.

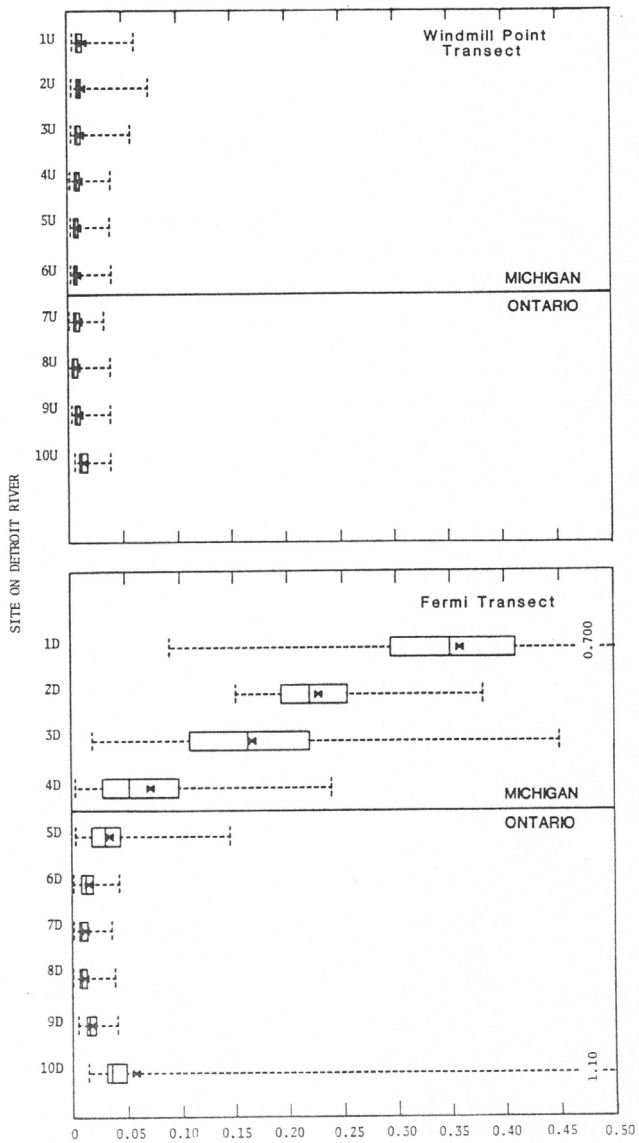


Figure 2.21 Boxplots of total ammonia nitrogen concentrations (mg/L as N) at two transects on the Detroit River (from Holtschlag, 1987)

2.2.5 Q-Q Plots

Direct comparisons can be made between two data sets by graphing the quantiles (percentiles) of one versus the quantiles (percentiles) of the second. This is called a quantile-quantile or Q-Q plot (Chambers et al., 1983). If the two data sets came from the same distribution, the quantile pairs would plot along a straight line with Yp = Xp, where p is the plotting position and Yp is the pth quantile of Y. In this case it would be said that the median, the quartiles, the 10th and 90th percentiles, etc., of the two data sets were equal. If one data set had the same shape as the second, differing only by an additive amount (each quantile was 5 units higher than for the other data set, for example), the quantile pairs would fall along a line parallel to but offset from the Yp=Xp line, also with slope =1. If the data sets differed by a multiplicative constant

(Yp = 5 • Xp, for example), the quantile pairs would lie along a straight line with slope equal to the multiplicative constant. More complex relationships will result in pairs of quantiles which do not lie along a straight line. The question of whether or not data sets differ by additive or multiplicative relationships will become important when hypothesis testing is conducted.

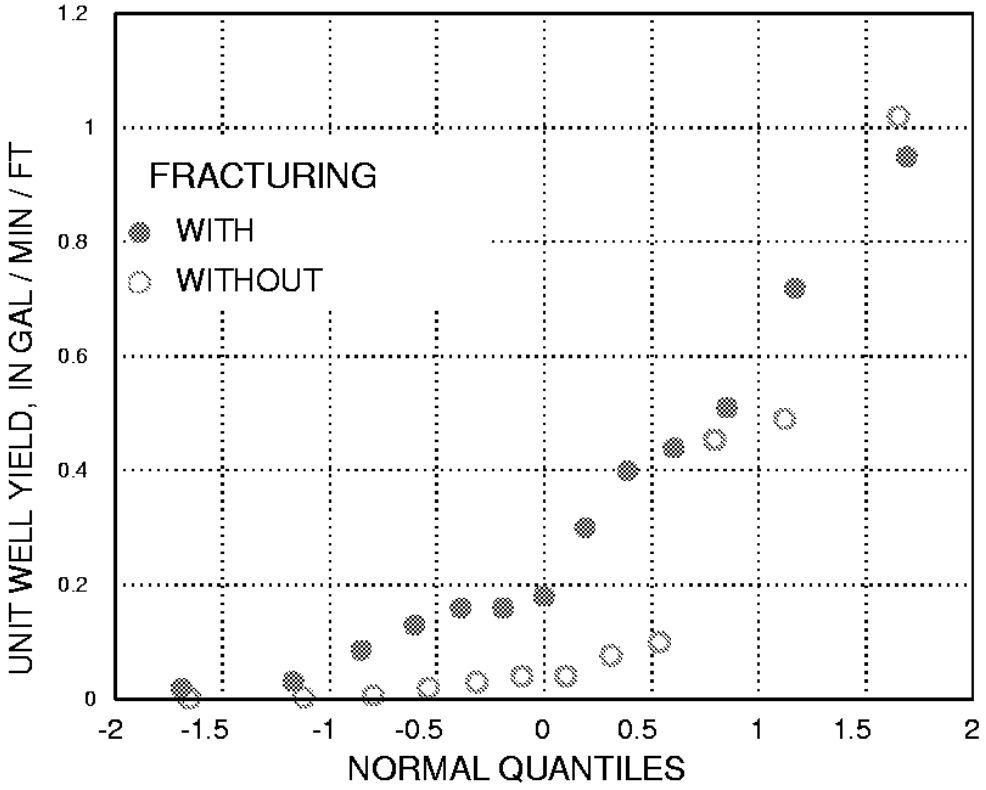


Figure 2.22 Probability plot of the unit well yield data

Figure 2.23 is a Q-Q plot of the well yield data. Several aspects of the relationship between the two data sets are immediately seen. First, the lowest 9 quantile pairs appear to fall along a straight line with a slope greater than 1, not parallel to the Yp = Xp line shown as a reference. This indicates a multiplicative relation between the data, with Y ≅ 4.4•X, where 4.4 is the slope of those data on the plot. Therefore, the yields with fractures are generally 4.4 times those without fractures for the lowest 75 percent of the data. The 3 highest quantile pairs return near to the Y = X line, indicating that the higher yields in the two data sets approach being equal. The hydrologist might be able to explain this phenomenon, such as higher yielding wells are deeper and less dependent on fracturing, or that some of the wells were misclassified, etc. Therefore the Q-Q plot becomes a valuable tool in understanding the relationships between data sets prior to performing any hypothesis tests.

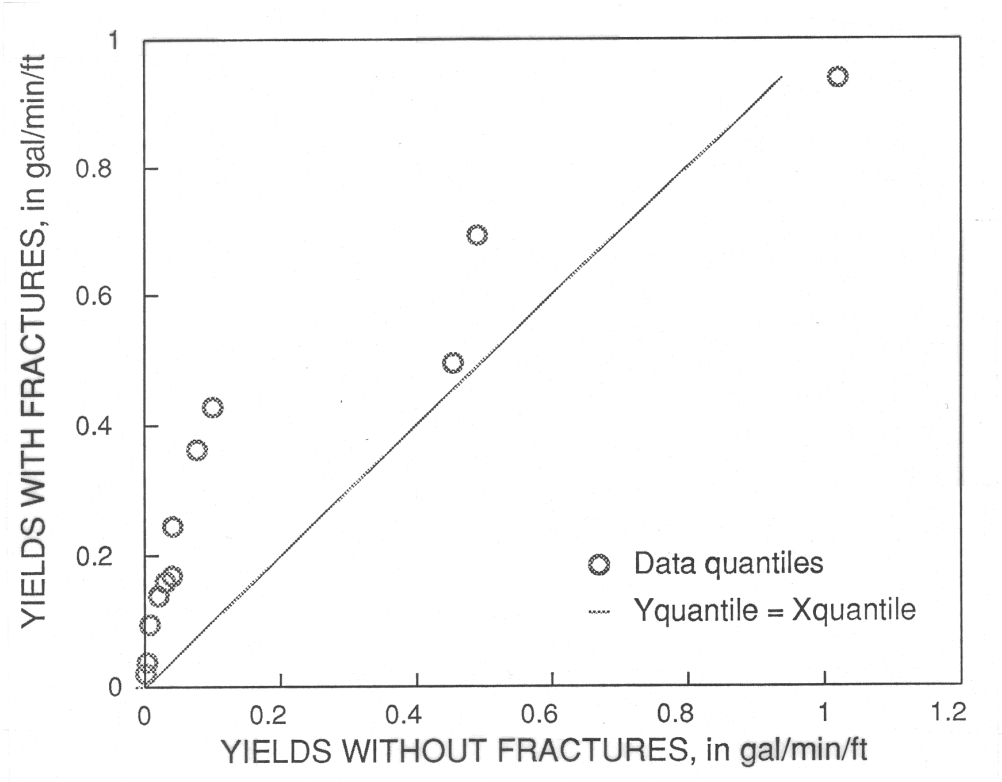


Figure 2.23 Q-Q plot of the unit well yield data

2.2.5.1 Construction of Q-Q plots

Q-Q plots are similar to probability plots. Now instead of plotting data quantiles from one group against quantiles of a theoretical distribution such as the normal, they are plotted against quantiles of a second data group.

When sample sizes of the two groups are identical, the x's and y's can be ranked separately, and the Q-Q plot is simply a scatterplot of the ordered data pairs $(x_1, y_1).....(x_n, y_n)$. When sample sizes are not equal, consider n to be the sample size of the smaller data set and m to be the sample size of the larger data set. The data values from the smaller data set are its pth quantiles, where $p = (i–0.4)/(n+0.2)$. The n corresponding quantiles for the larger data set are interpolated values which divide the larger data set into n equally-spaced parts. The following example illustrates the procedure.

For the well yield data, the 12 values without fractures designated x_i, i = 1,....n are themselves the sample quantiles for the smaller data set. Repeating the without fractures data given earlier in the chapter:

Unit well yields, in gal / min / ft (Wright, 1985)

x_i = yield without fractures p_i = plotting position y_i = yields with fractures

i	x_i	p_i	y_j	i	x_i	p_i	y_j	i	x_i	p_i	y_j
1	0.001	.05	-	5	0.030	.38	-	9	0.10	.70	-
2	0.003	.13	-	6	0.040	.46	-	10	0.454	.79	-
3	0.007	.21	-	7	0.041	.54	-	11	0.49	.87	-
4	0.020	.30	-	8	0.077	.62	-	12	1.02	.95	-

The .05 quantile (5th percentile) value of 0.001, for example, is to be paired on the Q-Q plot with the .05 quantile of the yields with fractures. To compute the corresponding y quantiles for the second data set, $p = (j - 0.4)/(m + 0.2)$, and therefore j must be:

$$\frac{(j - 0.4)}{(m + 0.2)} = \frac{(i - 0.4)}{(n + 0.2)} \text{ , or}$$

$$j = \frac{(m + 0.2) \bullet (i - 0.4)}{(n + 0.2)} + 0.4 \qquad [2.1]$$

If j is an integer, the data value y_j itself is plotted versus x_i. Usually, however, j will lie between two integers, and the y quantile must be linearly interpolated between the y data corresponding to the ranks on either side of j:

$$y_j = y_{j'} + (j - j') \bullet (y_{(j'+1)} - y_{j'}) \qquad [2.2]$$

where j' = integer (j)

For example, the well yield data with fractures are the following:
0.020 0.031 0.086 0.130 0.160 0.160 0.180 0.300 0.400 0.440 0.510 0.720 0.950 .

Therefore n = 12 m = 13 and from eq. 2.1, j = 1.08i – 0.03 .
The first of the 12 quantiles to be computed for the data with fractures is then:

i = 1 j =1 .05 j' = 1 $y_j = y_1 + 0.05 \bullet (y_2 - y_1)$

$0.020 + 0.05 \bullet (.031 - .020)$

$= 0.021$

All 12 quantiles are similarly interpolated:

i	j	interpolated y_j	i	j	y_j
1	1.05	0.021	7	7.53	0.245
2	2.13	0.038	8	8.61	0.362
3	3.21	0.095	9	9.69	0.428
4	4.29	0.139	10	10.77	0.495
5	5.37	0.160	11	11.85	0.692
6	6.45	0.169	12	12.93	0.939

These interpolated values are added to the table of quantiles given previously:

x_i = yields without fractures				p_i = plotting position				y_j = yields with fractures			
i	x_i	p_i	y_j	i	x_i	p_i	y_j	i	x_i	p_i	y_j
1	0.001	.05	0.021	5	.030	.38	0.160	9	0.10	.70	0.428
2	0.003	.13	0.038	6	.040	.46	0.169	10	0.454	.79	0.495
3	0.007	.21	0.095	7	.041	.54	0.245	11	0.49	.87	0.692
4	0.020	.30	0.139	8	.077	.62	0.362	12	1.02	.95	0.939

These (x_i, y_j) pairs are the circles which were plotted in figure 2.23.

2.3 Scatterplots and Enhancements

The two-dimensional scatterplot is one of the most familiar graphical methods for data analysis. It illustrates the relationship between two variables. Of usual interest is whether that relationship appears to be linear or curved, whether different groups of data lie in separate regions of the scatterplot, and whether the variability or spread is constant over the range of data. In each case, an enhancement called a "smooth" enables the viewer to resolve these issues with greater clarity than would be possible using the scatterplot alone. The following sections discuss these three uses of the scatterplot, and the enhancements available for each use.

2.3.1 Evaluating Linearity

Figure 2.24 is a scatterplot of the mass load of transported sand versus stream discharge for the Colorado R. at Lees Ferry, Colorado, during 1949-1964. Are these data sufficiently linear to fit a linear regression to them, or should some other term or transformation be included in order to account for curvature? In Chapters 9 and 11, other ways to answer this question will be presented, but many judgements on linearity are made solely on the basis of plots. To aid in this judgement, a "smooth" will be superimposed on the data.

The human eye is an excellent judge of the range of data on a scatterplot, but has a difficult time accurately judging the center -- the pattern of how y varies with x. This results in two difficulties

with judging linearity on a scatterplot as evident in figure 2.24. Outliers such as the two lowest sand concentrations may fool the observer into believing a linear model may not fit. Alternatively, true changes in slope are often difficult to discern from only a scatter of data To aid in seeing central patterns without being strongly influenced by outliers, a resistant center line can be fit to the data whose direction and slope varies locally in response to the data themselves. Many methods are available for constructing this type of center line -- probably the most familiar is the (non-resistant) moving average. All such methods may be called a "middle smooth", as they smooth out variations in the data into a coherent pattern through the middle. We discuss computation of smooths in Chapter 10. For now, we will merely illustrate their use as aids to graphical data analysis. The smoothing procedure we prefer is called LOWESS, or LOcally WEighted Scatterplot Smoothing (Cleveland and McGill, 1984b; Cleveland, 1985).

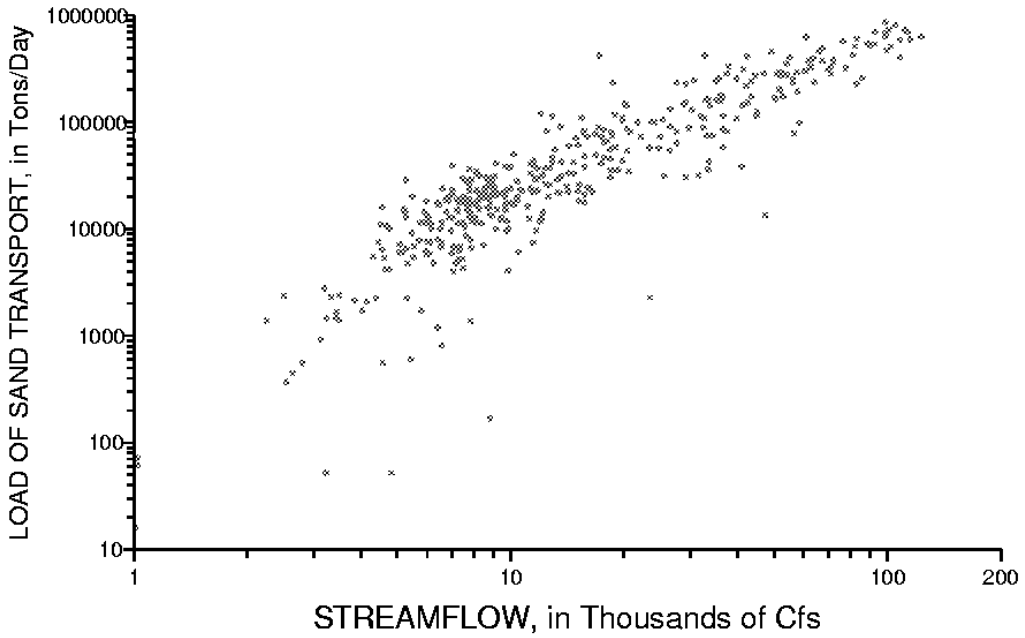


Figure 2.24 Suspended sand transport at Lees Ferry, Arizona, 1949-1952

Figure 2.25 presents the Lees Ferry sediment data of figure 2.24, with a superimposed middle smooth. Note the nonlinearity now evident by the curving smooth on the left-hand side of the plot. The rate of sand transport slows above 6600 ($e^{8.8}$) cfs. This curvature is easier to see with the superimposed smooth. It is important to remember that no model, such as a linear or quadratic function, is assumed prior to computing a smooth. The smoothed pattern is totally derived by the pattern of the data, and may take on any shape. As such, smooths are an exploratory tool for discerning the form of relationship between y and x. Seeing the pattern of figure 2.25, a quadratic term might be added, a piecewise linear fit used, or a transformation stronger than logs used prior to performing a linear regression of concentration versus discharge (see Chapter 9).

Middle smooths should be regularly used when analyzing data on scatterplots, and when presenting those data to others. As no model form is assumed by them, they let the data describe the pattern of dependence of y on x. Smooths are especially useful when large amounts of data are to be plotted, and several groups of data are placed on the same plot. For example, Welch and others (1988) depicted the dependence of log of arsenic concentration on pH for thousands of groundwater samples throughout the western United States (figure 2.26). By using middle smooths, data from one physiographic province was seen to differ from the other three provinces in its relationship between pH and arsenic.

2.3.2 Evaluating Differences in Location on a Scatterplot

Figure 2.27 is a scatterplot of conductance versus pH for samples collected at low-flow in small streams within the coal mining region of Ohio (data from Helsel, 1983). Each stream was classified by the type of land it was draining -- unmined land, lands mined and later reclaimed, and lands mined and then abandoned without reclamation. These three types of upstream lands are plotted with different symbols in figure 2.27.

To see the three locations more clearly, a smooth can be constructed for each group which encloses either 50 or 75 percent of the data. This type of smooth is called a polar smooth (Cleveland and McGill, 1984b), and its computation is detailed in Chapter 10. Briefly, the data are transformed into polar coordinates, a middle or similar smooth computed, and the smooth is re-transformed back into the original units. In figure 2.28. a polar smooth enclosing 75 percent of the data in each of the types of upstream land is plotted. These smooths are again not limited to a prior shape or form, such as that of an ellipse. Their shapes are determined from the data.

Polar smooths can be a great aid in exploratory data analysis. For example, the irregular pattern for the polar smooth of data from abandoned lands in figure 2.28 suggests that two separate subgroups are present, one with higher pH than the other. Using different symbols for data from each of the two geologic units underlying these streams shows indeed that the basins underlain by a limestone unit have generally higher pH than those underlain by a sandstone. Therefore the type of geologic unit should be included in any analysis or model of the behavior of chemical constituents for these data.

Polar smooths are especially helpful when there is a large amount of data to be plotted on a scatterplot. In such situations, the use of different symbols for distinguishing between groups will be ineffective, as the plot will be too crowded to see patterns in the locations of symbols. Indeed, in some locations it will not be possible to distinguish which symbol is plotted. Plots presenting small data points and the polar smooths as in figure 2.28, or even just the polar smooths themselves, will provide far greater visual differentiation between groups.

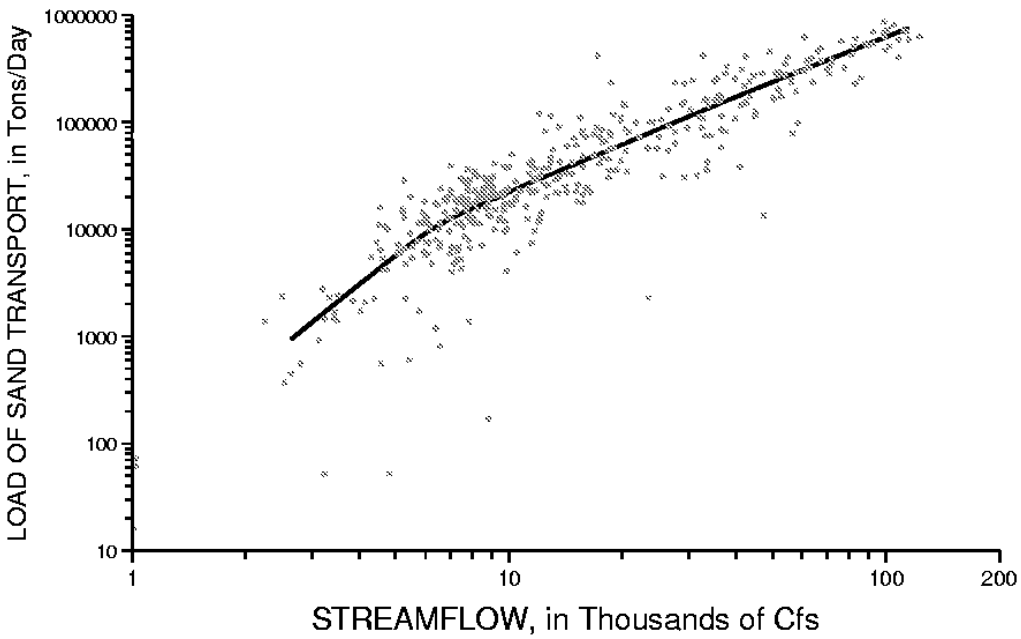


Figure 2.25 Data of figure 2.24 with superimposed lowess smooth

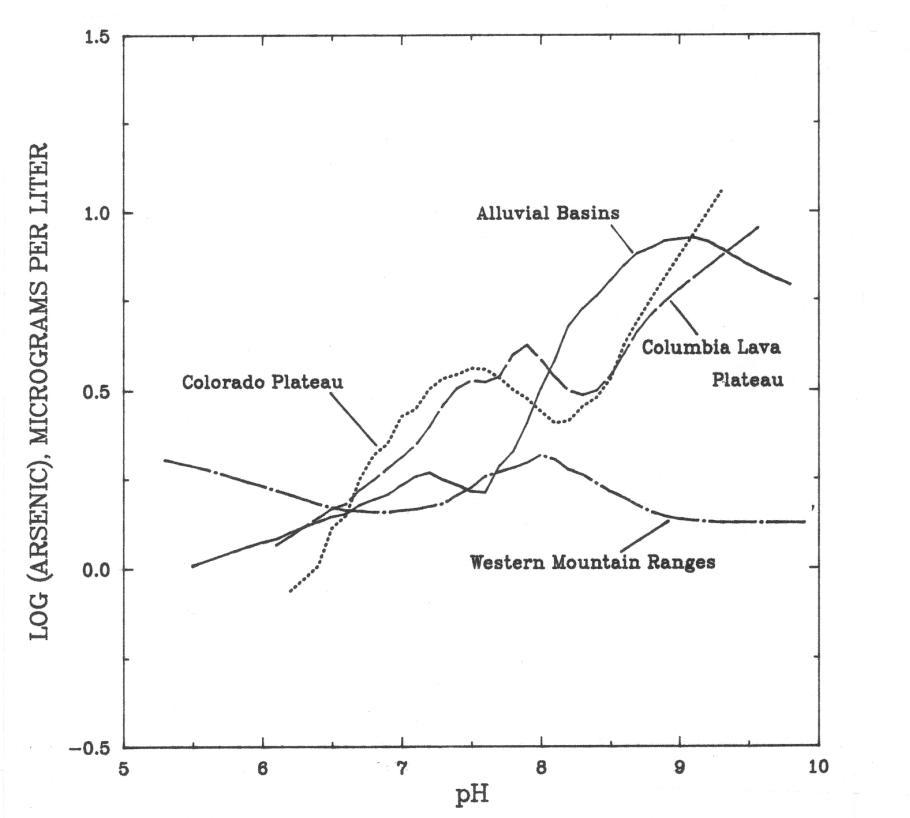


Figure 2.26 Dependence of log(As) on pH for 4 areas in the western U.S. (Welch and others, 1988)

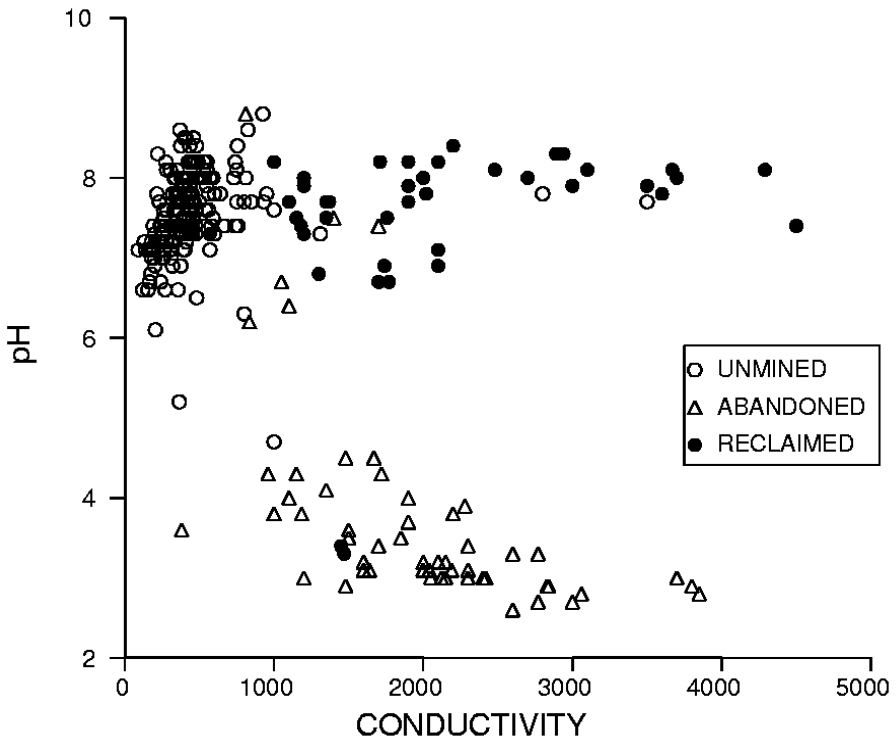


Figure 2.27 Scatterplot of water-quality draining three types of upstream land use

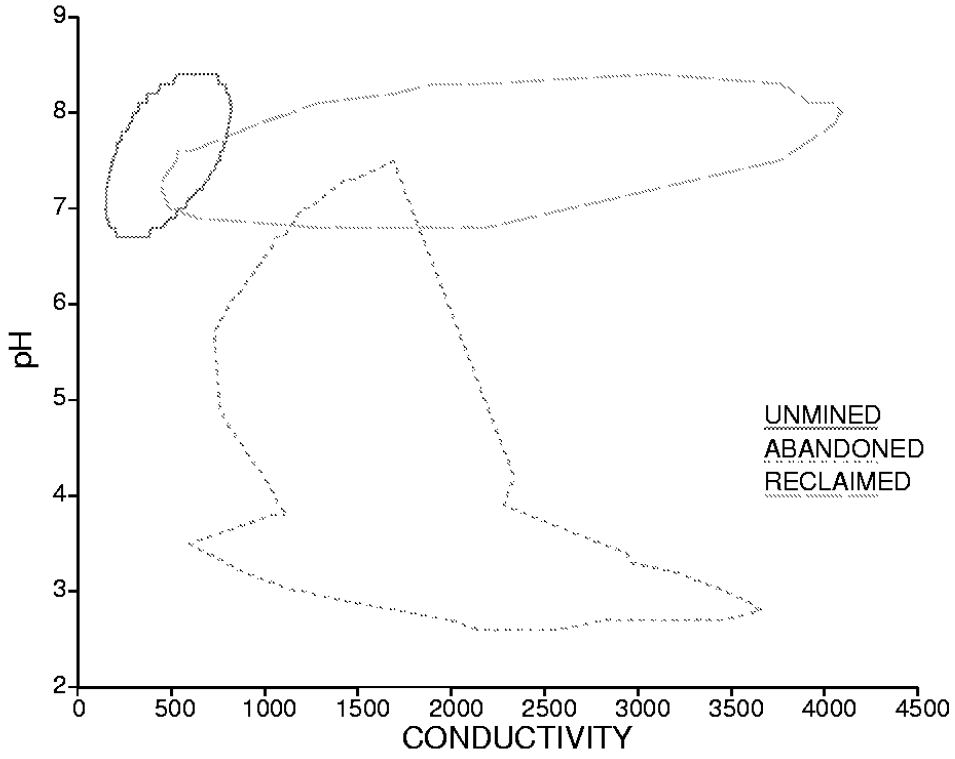


Figure 2.28 Polar smooths for the three groups of data in figure 2.27

2.3.3 Evaluating Differences in Spread

In addition to understanding where the middle of data lie on a scatterplot, it is often of interest to know something about the spread of the data as well. Homoscedasticity (constant variance) is a crucial assumption of ordinary least-squares regression, as we will see later. Changes in variance also invalidate parametric hypothesis test procedures such as analysis of variance. From a more exploratory point of view, changes in variance may be as important or more important than changes in central value. Differences between estimation methods for flood quantiles, or between methods of laboratory analysis of some chemical constituent, are often differences in repeatability of the results and not of method bias. Graphs again can aid in judging differences in data variability, and are often used for this purpose.

A major problem with judgements of changing spread on a scatterplot is again that the eye is sensitive to seeing the range of data. The presence of a few unusual values may therefore incorrectly trigger a perception of changing spread. This is especially a problem when the density of data changes across a scatterplot, a common occurrence. Assuming the distribution of data to be identical across a scatterplot, and that no changes in variablility or spread actually occur, areas where data are more dense are more likely to contain outlying values on the plot, and the range of values is likely to be larger. This leads to a perception that the spread has changed.

One graphical means of determining changes in spread has been given by Chambers et al. (1983). First, a middle smooth is computed, as in figure 2.25. The absolute values of differences $|d_i|$ between each data point and the smooth at its value of x is a measure of spread.

$$|d_i| = |y_i - l_i| \qquad \text{where } l_i \text{ is the value for the lowess smooth at } x_i \qquad [2.3]$$

By graphing these absolute differences $|d_i|$ versus x_i, changes in spread will show as changes in absolute differences. A middle smooth of these differences should also be added to make the pattern more clear. This is done in figure 2.29, a plot of the absolute differences between sand concentration and its lowess smooth for the Lees Ferry data of figure 2.25. Note that there is a slight decrease in $|d_i|$, indicating a small decrease of variability or spread in concentration with increasing discharge at that site.

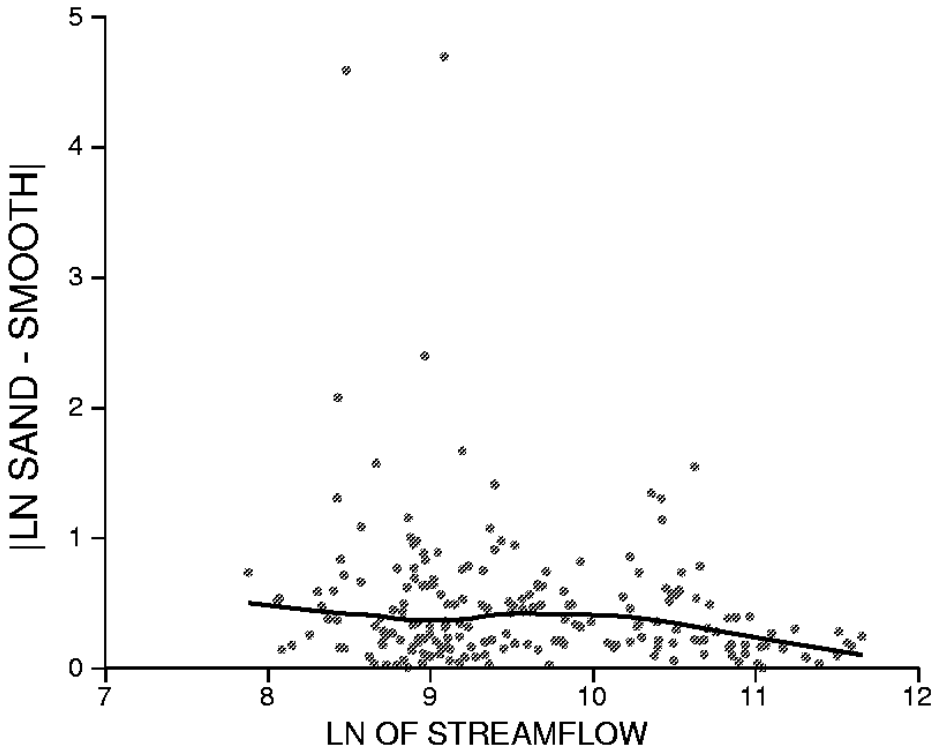


Figure 2.29 Absolute residuals show whether the spread changes with changing x -- sediment concentrations at Lees Ferry, Arizona

2.4 Graphs for Multivariate Data

Boxplots effectively illustrate the characteristics of data for a single variable, and accentuate outliers for further inspection. Scatterplots effectively illustrate the relationships between two variables, and accentuate points which appear unusual in their x-y relationship. Yet there are numerous situations where relationships between more than two variables should be considered simultaneously. Similarities and differences between groups of observations based on 3 or more variables are frequently of interest. Also of interest is the detection of outliers for data with multiple variables. Graphical methods again can provide insight into these relationships. They supplement and enhance the understanding provided by formal hypothesis test procedures. Two multivariate graphical methods already are widely used in water-quality studies -- Stiff and Piper diagrams. These and other graphical methods are outlined in the following sections. For more detailed discussions on multivariate graphical methods, see Chambers et al. (1983), or the textbook by Everitt (1978).

2.4.1 Profile Plots

Profile plots are a class of graphical methods which assign each variable to a separate and parallel axis. One observation is represented by a series of points, one per axis, which are

connected by a straight line forming the profile. Each axis is scaled independently, based on the range of values in the entire data set. Comparisons between observations are made by comparing profiles.

As an example, assume that sediment loads are to be regionalized. That is, mean annual loads are to be predicted at ungaged sites based on basin characteristics (physical and climatic conditions) at those sites. Of interest may be the interrelationships between sites based on their basin characteristics, as well as which characteristics are associated with high or low annual values. Profile plots such as the one of site basin characteristics in figure 2.30 would effectively illustrate those relationships.

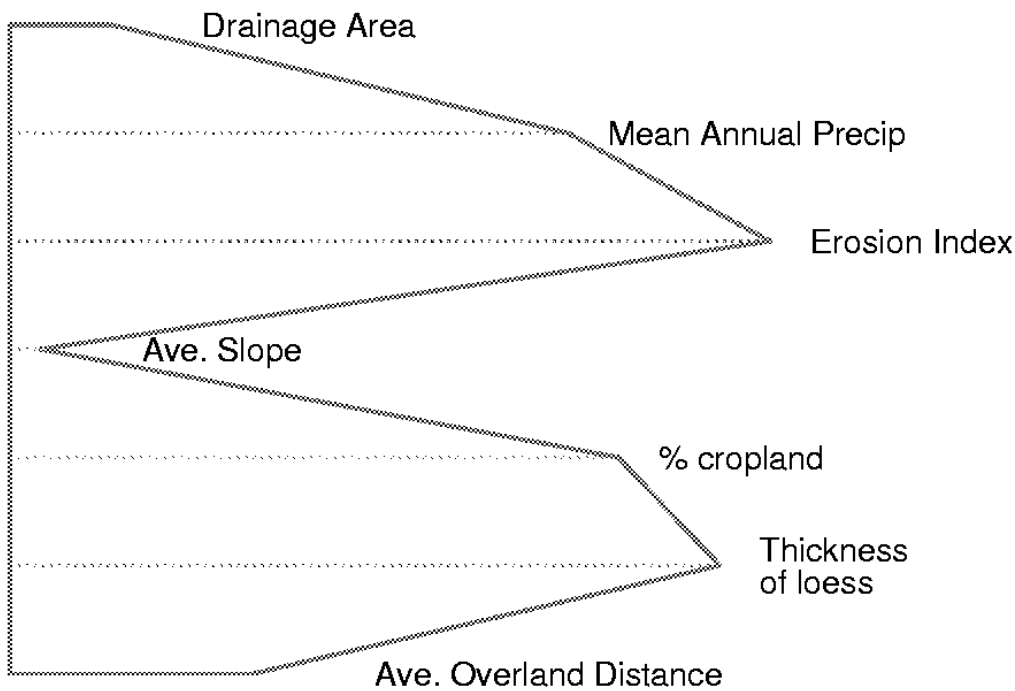


Figure 2.30 Profile plot of selected basin characteristcs, Cow Creek near Lyons, Kansas (data from Jordan, 1979).

2.4.1.1 Stiff diagrams

Stiff diagrams (Hem, 1985) are the most familiar application of profile plots in water resources. In a Stiff diagram, the milliequivalents of major water-quality constituents are plotted for a single sample, with the cation profile plotted to the left of the center line, and anion profile to the right (figure 2.31). Comparisons between several samples based on multiple water-quality constituents is then easily done by comparing shapes of the Stiff diagrams. Figure 2.32 shows one such comparison for 14 groundwater samples from the Fox Hills Sandstone in Wyoming (Henderson, 1985).

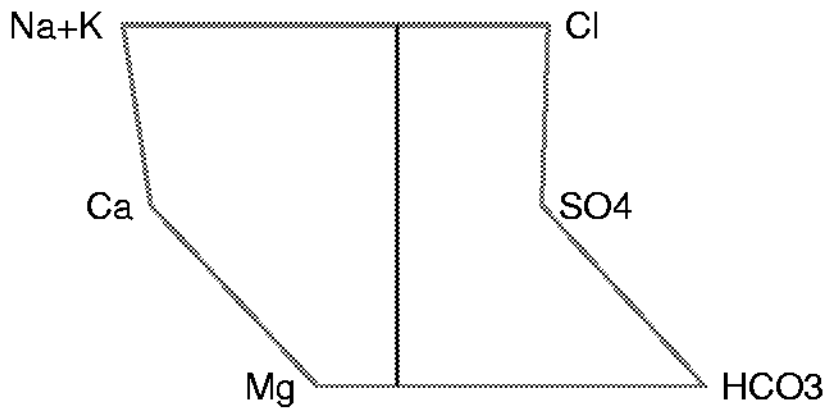


Figure 2.31 Stiff diagram for a groundwater sample from the Columbia River Basalt aquifer, Oregon (data from Miller and Gonthier, 1984).

2.4.2 Star Plots

A second method of displaying multiple axes is to have them radiate from a central point, rather than aligned parallel as in a profile plot. Again, one observation would be represented by a point on each axis, and these points are connected by line segments. The resulting figures resemble a star pattern, and are often called star plots. Angles between rays of the star are $360^{o}/k$, where k is the number of axes to be plotted. To provide the greatest visual discrimination between observations, rays measuring related characteristics should be grouped together. Unusual observations will stand out as a star looking quite different than the other data, perhaps having an unusually long or short ray. In figure 2.33, the basalt water-quality data graphed using a Stiff diagram in figure 2.31 is displayed as a star plot. Note that the cations are grouped together on the top half of the star, with anions along the bottom.

2.4.2.1 Kite diagrams

A simplified 4-axis star diagram, the "kite diagram", has been used for displaying water-quality compositions, especially to portray compositions of samples located on a map (Colby, 1956). Cations are plotted on the two vertical axes, and anions on the two horizontal axes. The primary advantage of this plot is its simplicity. Its major disadvantage is also its simplicity, in that the use of only four axes may hide important characteristics of the data. One might need to know whether calcium or magnesium were present in large amounts, for example, but that could not be determined from the kite diagram. There is no reason why a larger number of axes could not be employed to give more detail, making the plot a true star diagram. Compare for example the basalt data plotted as a star diagram in figure 2.33 and as a kite diagram in figure 2.34.

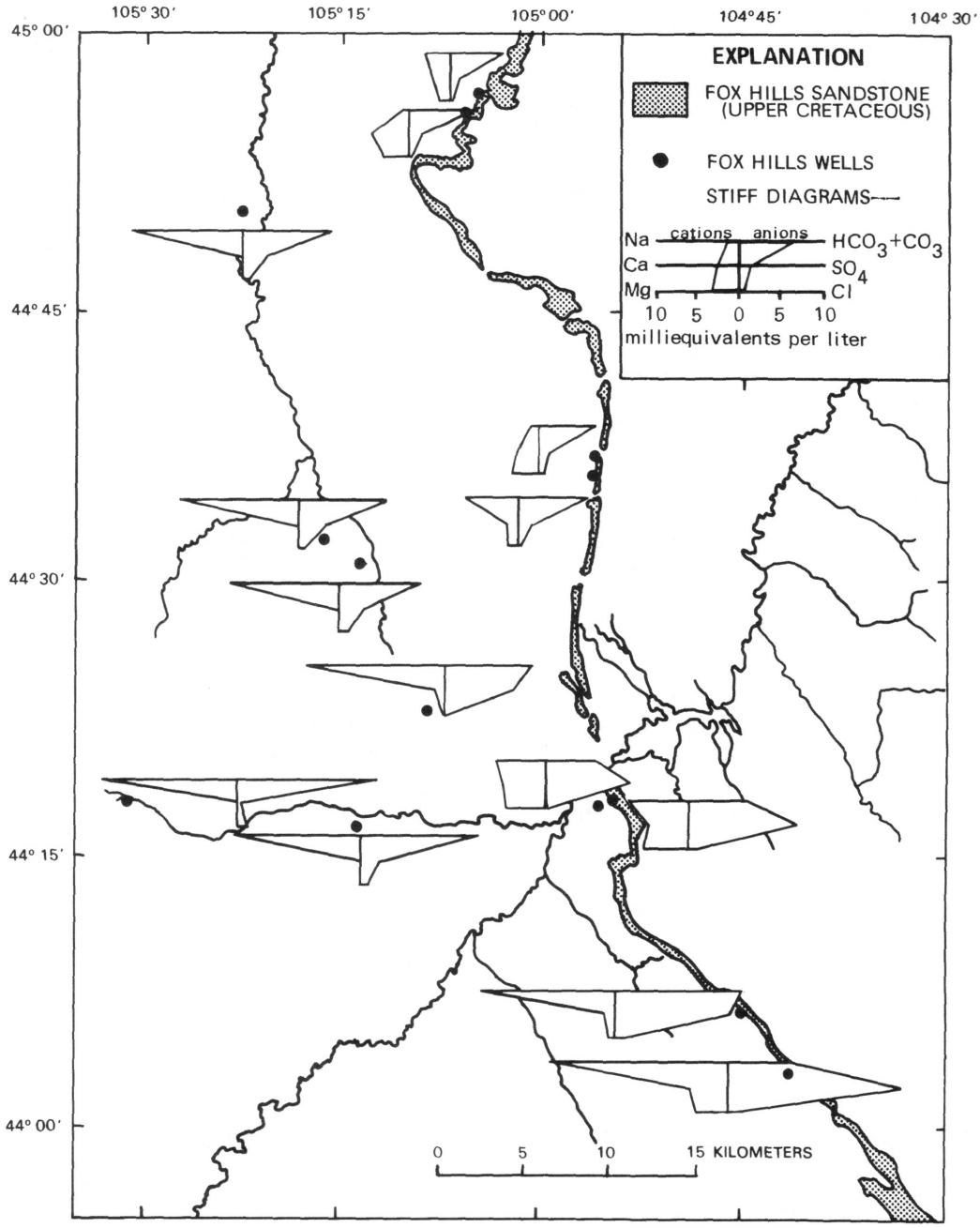


Figure 2.32 Stiff diagrams to display areal differences in water quality (from Henderson, 1985)

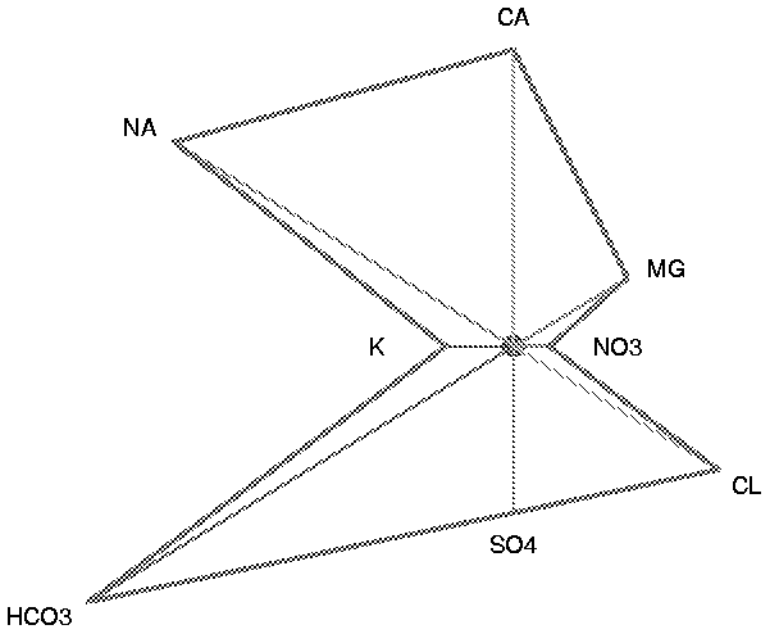


Figure 2.33 Star diagram of the basalt water-quality data

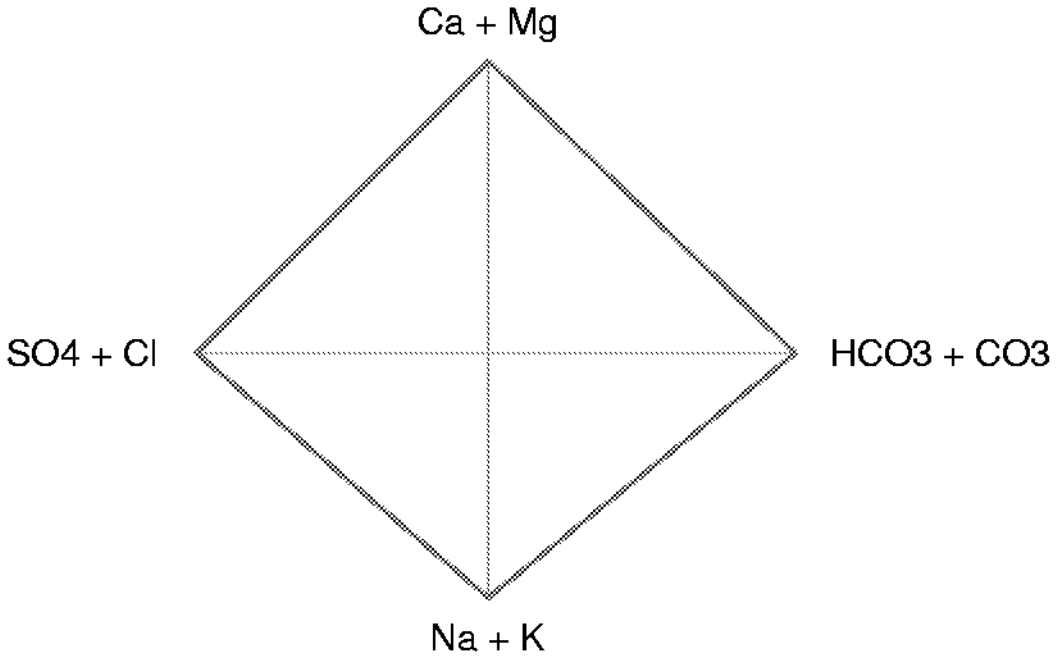


Figure 2.34 Kite diagram of the basalt water-quality data

One innovative use of the kite diagram was made by Davis and Rogers (1984). They plotted the quartiles of all observations taken from each of several formations, and at different depth ranges, in order to compare water quality between formations and depths (figure 2.35). The kite plots in

this case are somewhat like multivariate boxplots. There is no reason why the other multivariate plots described here could not also present percentile values for a group of observations rather than descriptions of individual values, and be used to compare among groups of data.

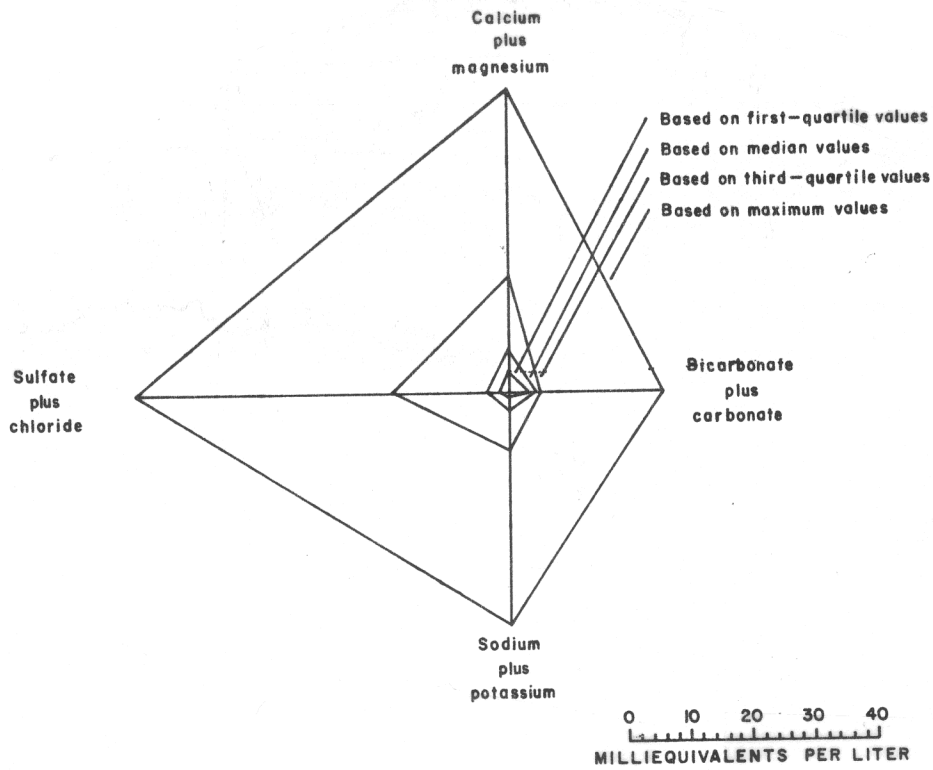


Figure 2.35 Kite diagram of quartiles of composition from an alluvial formation in Montana (from Davis and Rogers, 1984).

2.4.3 Trilinear Diagrams

Trilinear diagrams have been used within the geosciences since the early 1900's. When three variables for a single observation sum to 100 percent, they can be represented as one point on a triangular (trilinear) diagram. Figure 2.36 is one example -- three major cation axes upon which is plotted the cation composition for the basalt data of figure 2.31. Each of the three cation values, in milliequivalents, is divided by the sum of the three values, to produce a new scale in percent of total cations:

$$c_i = m_i / (m_1 + m_2 + m_3) \qquad [2.4]$$

where the c_i is in percent of total cations, and m_i are the milliequivalents of cation i.

For the basalt data, Ca = 0.80 meq, Mg = 0.26 meq, and Na+K = 0.89 meg. Thus %Ca = 41, %Mg = 13, and %[Na + K] = 46. As points on these axes sum to 100 percent, only two of the variables are independent. By knowing two values c_1 and c_2, the third is also known: $c_3 = (100 - c_1 - c_2)$.

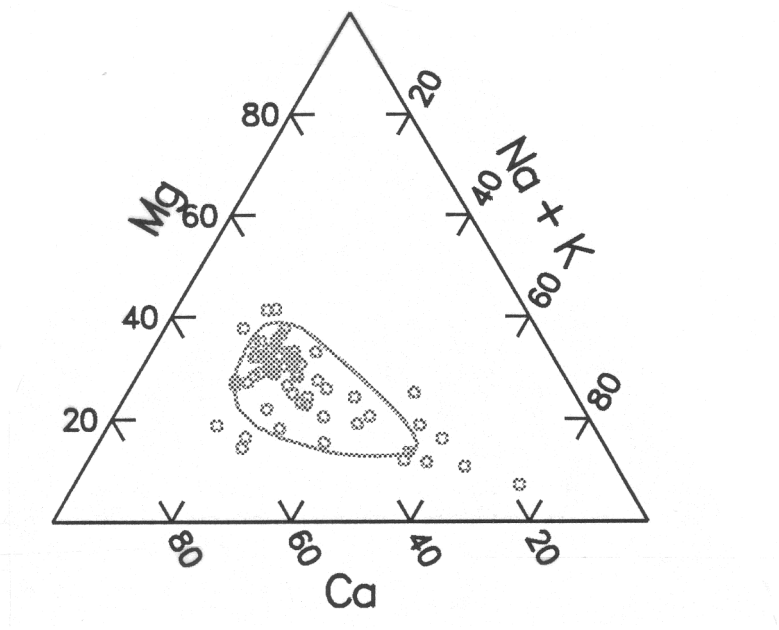


Figure 2.36 Trilinear diagram for the basalt cation composition (units are percent milliequivalents).

2.4.3.1 Piper diagrams

Piper (1944) applied these trilinear diagrams to both cation and anion compositions of water qualtiy data. He also combined both trilinear diagrams into a single summary diagram with the shape of a diamond (figure 2.37). This diamond has four sides, two for cations and two for anions. However, it also has only two independent axes, one for a cation (say Ca + Mg), and one for an anion (say Cl + SO_4). If the (Ca + Mg) percentage is known, so is the (Na + K) percentage, as one is 100% minus the other, and similarly for the anions. The collection of these three diagrams in the format shown in figure 2.37 is called a Piper diagram.

Piper diagrams have the advantage over Stiff and star diagrams that each observation is shown as only one point. Therefore, similarities and differences in composition between numerous observations is more easily seen with Piper diagrams. Stiff and star diagrams have two advantages over Piper diagrams: 1) they may be separated in space and placed on a map or other graph, and 2) more than four independent attributes (two cation and two anion) can be displayed at one time. Thus the choice of which to use will depend on the purpose to which they are put.

Envelopes have been traditionally drawn by eye around a collection of points on a Piper diagram to describe waters of "similar" composition. Trends (along a flow path, for example) have traditionally been indicated by using different symbols on the diagram for different data groups, such as for upgradient and downgradient observations, and drawing an arrow from one group to

the other. Recently, both of these practices have been quantified into significance tests for differences and trends associated with Piper diagrams (Helsel, 1992). Objective methods for drawing envelopes (a smoothed curve) and trend lines on a Piper diagram were also developed. The envelope drawn on figure 2.37 is one example. Smoothing procedures are discussed in more detail in Chapter 10.

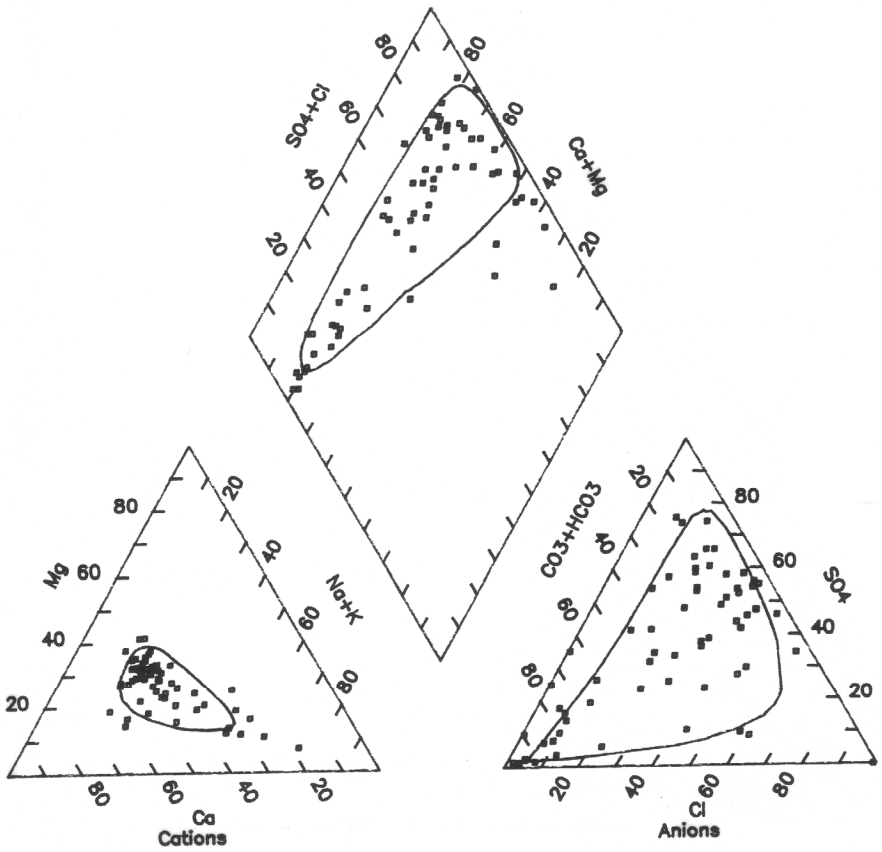


Figure 2.37 Piper diagram of groundwaters from the Columbia River Basalt aquifer in Oregon (data from Miller and Gonthier, 1984)

2.4.4 Plots of Principal Components

One method for viewing observations on multiple axes is to reduce the number of axes to two, and then plot the data as a scatterplot. An important dimension reduction technique is principal components analysis, or PCA (Johnson and Wischern, 1982).

Principal components are linear combinations of the p original variables which form a new set of variables or axes. These new axes are uncorrelated with one another, and have the property that the first principal component is the axis that explains more of the variance of the data than any other axis. The second principal component explains more of the remaining variance than any other axis which is uncorrelated with (orthogonal to) the first. The resulting p axes are thus

new "variables", the first few of which often explain the major patterns of the data in multivariate space. The remaining principal components may be treated as residuals, measuring the "lack of fit" of observations along the first few axes.

Each observation can be located on the new set of principal component (pc) axes. For example, suppose principal components were computed for four original variables, the cations Ca, Mg, Na and K. The new axes would be linear combinations of these variables, such as:

pc1 = 0.75 Ca + 0.8 Mg +0.1 Na +0.06 K	a "calcareous" axis ?
pc2 = 0.17 Ca + 0.06 Mg +0.6 Na + 0.8 K	a "Na + K" axis?
pc3 = 0.4 Ca – 0.25 Mg – 0.1 Na + 0.1 K	a "Ca vs. Mg" axis?
pc4 = 0.05 Ca – 0.1 Mg + 0.1 Na + 0.2 K	residual noise

An observation which had milliequivalents of Ca = 1.6, Mg = 1.0, Na = 1.3 and K = 0.1 would have a value on pc1 equal to (0.6 •1.6 + 0.8 •1.0 + 0.1 •1.3 + 0.06 •0.1) = 1.9, and similarly for the other new "variables". At this point no reduction in dimensions has taken place, as each observation still has values along the p=4 new pc axes, as they did for the 4 original axes.

Now, however, plots can be made of the locations of observations oriented along the new principal components axes. Most notably, a scatterplot for the first two components (pc1 vs. pc2) will show how the observations group together along the new axes which now contain the most important information about the variation in the data. Thus groupings in multivariate space have been simplified into groupings along the two most important axes, allowing those groupings to be seen by the data analyst. Waters with generally different chemical compositions should plot at different locations on the pc scatterplot. Data known to come from two different groups may be compared using boxplots, probability plots, or Q-Q plots, but now using the first several pc axes as the measurement "variables". Additionally, plots can be made of the <u>last</u> few pc axes, to check for outliers. These outliers in multivariate space will now be visible by using the "lack of fit" principal components to focus attention at the appropriate viewing angle. Outliers having unusually large or small values on these plots should be checked for measurement errors, unusual circumstances, and the other investigations outliers warrant. Examples of the use of plots of components include Xhoffer et al. (1991), Meglen and Sistko (1985), and Lins (1985).

2.4.5 Other Multivariate Plots

2.4.5.1 3-Dimensional rotation

If three variables are all that are under consideration, several microcomputer packages now will plot data in pseudo-3 dimensions, and allow observations to be rotated in space along all three axes. In this way the inter-relationships between the three variables can be visually observed, data visually clustered into groups of similar observations, and outliers discerned. In figure 2.38

two of the many possible orientations for viewing a data set were output from MacSpin (Donoho et al., 1985), a program for the Apple Macintosh. The data are water qualtity variables measured at low flow in basins with and without coal mining and reclamation (Helsel, 1983)

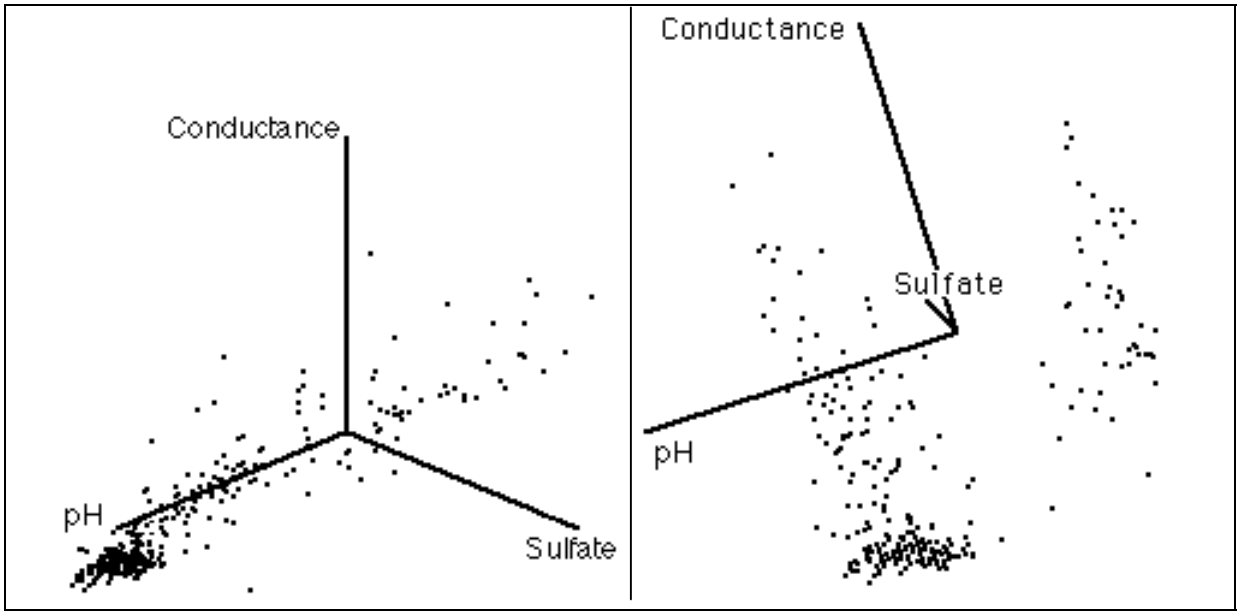


Figure 2.38 Two 3-dimensional plots of a water-quality data set

Note the u-shaped pattern in the data seen in the right-hand plot. There is some suggestion of two separate groups of data, the causes of which can be checked by the analyst. This pattern is not evident in the left-hand orientation. By rotating data around their three axes, patterns may be seen which would not be evident without a 3-dimensional perspective, and greater insight into the data is obtained.

2.4.5.2 Scatterplot matrix

Another method for inspecting data measured by p variables is to produce a scatterplot for each of the $p \bullet (p-1)/2$ possible pairs of variables. These are then printed all on one screen or page. Obviously, little detail can be discerned on any single plot within the matrix, but variables which are related can be grouped, linear versus nonlinear relationships discerned, etc. Chambers et al. (1983) describe the production and utility of scatterplot matrices in detail.

Figure 2.39 is a scatterplot matrix for 5 water-quality variables at low-flow from the coal mining data of Helsel (1983). On the lowest row are histograms for each individual variable. Note the right skewness for all variables except pH. All rows above the last contain scatterplots between each pair of varables. For example, the single plot in the first row is the scatterplot of conductance (cond) versus pH. Note the two separate subgroups of data, representing low and high pH waters. Evident from other plots are the linear association between conductance and

sulfate (SO4), the presence of high total iron concentrations (TFe) for waters of low alkalinity (ALK) and pH, and high TFe for waters of high sulfate and conductance.

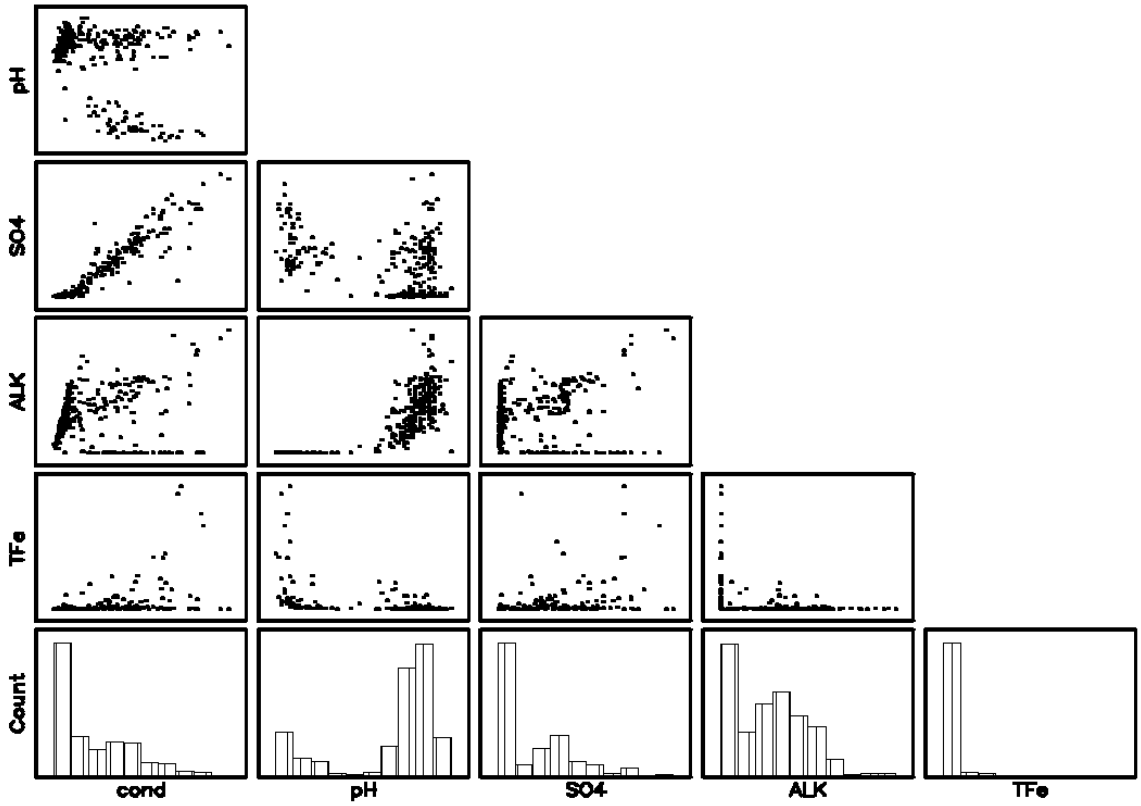


Figure 2.39 Scatterplot matrix showing the relationships between 5 water-quality variables

2.4.5.3 Methods to avoid

Two commonly-used methods should usually be avoided, as they provide little ability to compare differences between groups of data. These are stacked bar charts and pie charts. Both allow only coarse discrimination to be made between segments of the plot. Figure 2.40, for example, is a stacked bar chart of the basalt water-quality data previously shown as a Stiff (figure 2.31) and star (figure 2.33) plot. Note that only large differences between categories within a bar are capable of being discerned. For example, it is much easier to see that chloride (Cl) is larger than sulfate (SO4) on the Stiff diagram than on the stacked bar chart. In addition, stacked bar charts provide much less visual distinction when comparing differences among many sites, as in figure 2.32. Stiff or star diagrams allow differences to be seen as differences in shape, while stacked bar charts require judgements of length without a common datum, a very difficult type of judgement. Multiple pie charts require similarly imprecise and difficult judgements. Further information on these and other types of presentation graphics is given in the last chapter.

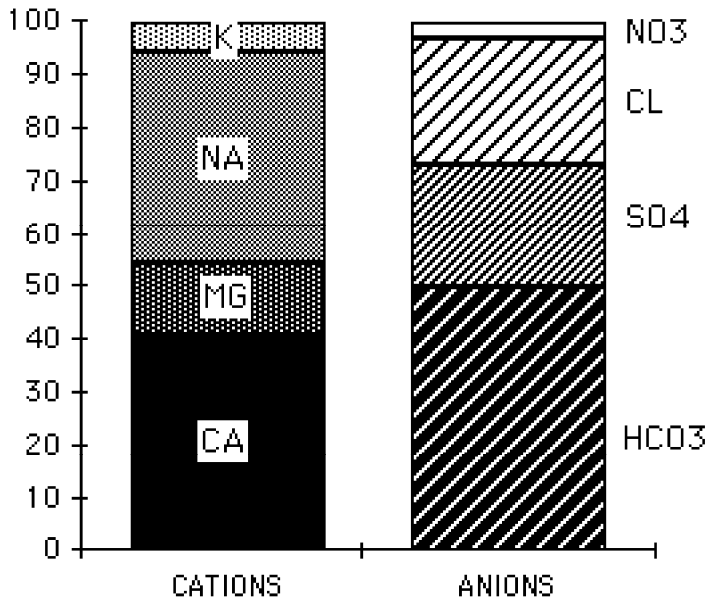


Figure 2.40 Stacked bar chart of the basalt data

Exercises

2.1 Annual peak discharges for the Saddle River in New Jersey are given in Appendix C1. For the peaks occuring from 1968-1989, draw

a) a histogram
b) a boxplot
c) a quantile plot (using $(i - .4)/(n + .2)$)

What transformation, if any, would make these data more symmetric?

2.2 Arsenic concentrations (in ppb) were reported for ground waters of southeastern New Hampshire (Boudette and others, 1985). For these data, compute

a) a boxplot
b) a probability plot

Based on the probability plot, describe the shape of the data distribution. What transformation, if any, would make these data more symmetric?

1.3	1.5	1.8	2.6	2.8	3.5	4.0	4.8
8	9.5	12	14	19	23	41	80
100	110	120	190	240	250	300	340
580							

2.3 Feth et al. (1964) measured chemical compositions of waters in springs draining differing rock types. Compare chloride concentrations from two of these rock types using a Q-Q plot. Also plot two other types of graphs. Describe the similarities and differences in chloride. What characteristics are evident in each graph?

	Chloride concentration, in mg/L					
Granodiorite	6.0	0.5	0.4	0.7	0.8	6.0
	5.0	0.6	1.2	0.3	0.2	0.5
	0.5	10	0.2	0.2	1.7	3.0
Qtz Monzonite	1.0	0.2	1.2	1.0	0.3	0.1
	0.1	0.4	3.2	0.3	0.4	1.8
	0.9	0.1	0.2	0.3	0.5	

2.4 The following chemical and biological data were reported by Frenzel (1988) above and below a waste treatment plant (WTP). Graph and compare the two sets of multivariate data. What effects has the WTP appeared to have?

	Above	Below	units
Chironomidae	2500	3200	ave # per substrate
Simuliidae	3300	230	ave # per substrate
Baetidae	2700	2700	ave # per substrate
Hydropsychidae	440	88	ave # per substrate
Native trout	6.9	7.9	# per 10,760 sq. ft.
Whitefish	140	100	# per 10,760 sq. ft.
Nongame fish	54	180	# per 10,760 sq. ft.
Aluminum in clays	1950	1160	μg/g
Organic Carbon	4.2	2.1	g/kg
Ammonia	0.42	0.31	mg/L as N

Chapter 3

Describing Uncertainty

The mean nitrate concentration in a shallow aquifer under agricultural land was calculated as 5.1 mg/L. How reliable is this estimate? Is 5.1 mg/L in violation of a health advisory limit of 5 mg/L? Should it be treated differently than another aquifer having a mean concentration of 4.8 mg/L?

Thirty wells over a 5-county area were found to have a mean specific capacity of 1 gallon per minute per foot, and a standard deviation of 7 gallons per minute per foot. A new well was drilled and developed with an acid treatment. The well produced a specific capacity of 15 gallons per minute per foot. To determine whether this increase might be due to the acid treatment, how likely is a specific capacity of 15 to result from the regional distribution of the other 30 wells?

An estimate of the 100-year flood, the 99th percentile of annual flood peaks, was determined to be 10,000 cubic feet per second (cfs). Assuming that the choice of a particular distribution to model these floods (Log Pearson Type III) is correct, what is the reliability of this estimate?

In chapter 1 several summary statistics were presented which described key attributes of a data set. They were sample estimates (such as $\bar{x}$ and s^2) of true and unknown population parameters (such as μ and σ^2). In this chapter, descriptions of the uncertainty or reliability of sample estimates is presented. As an alternative to reporting a single estimate, the utility of reporting a range of values called an "interval estimate" is demonstrated. Both parametric and nonparametric interval estimates are presented. These intervals can also be used to test whether the population parameter is significantly different from some pre-specified value.

3.1 Definition of Interval Estimates

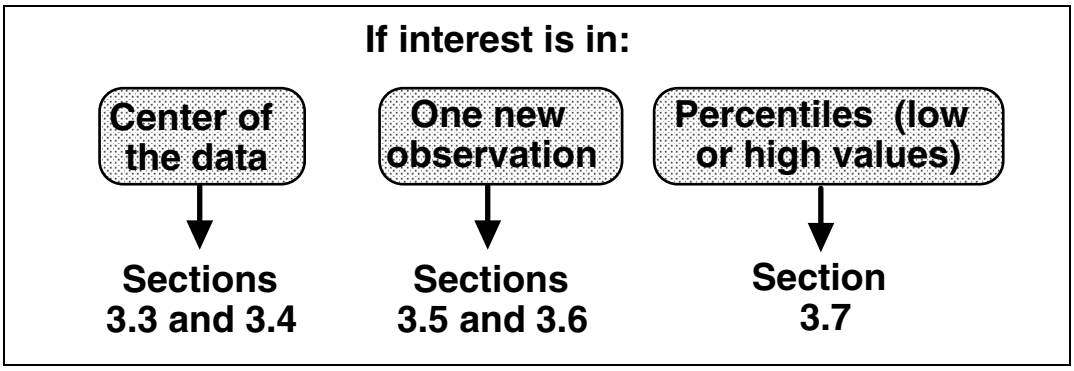
The sample median and sample mean estimate the corresponding center points of a population. Such estimates are called **point estimates**. By themselves, point estimates do not portray the reliability, or lack of reliability (variability), of these estimates. For example, suppose that two data sets X and Y exist, both with a sample mean of 5 and containing the same number of data. The Y data all cluster tightly around 5, while the X data are much more variable. The point estimate of 5 for X is much less reliable than that for Y because of the greater variability in the X data. In other words, more caution is needed when stating that 5 estimates the true population mean of X than when stating this for Y. Reporting only the sample (point) estimate of 5 fails to give any hint of this difference.

As an alternative to point estimates, **interval estimates** are intervals which have a stated probability of containing the true population value. The intervals are wider for data sets having greater variability. Thus in the above example an interval between 4.7 and 5.3 may have a 95% probability of containing the (unknown) true population mean of Y. It would take a much wider interval, say between 2.0 and 8.0, to have the same probability of containing the true mean of X. The difference in the reliability of the two estimates is therefore clearly stated using interval estimates. Interval estimates can provide two pieces of information which point estimates cannot:

1. A statement of the probability or likelihood that the interval contains the true population value (its reliability).
2. A statement of the likelihood that a single data point with specified magnitude comes from the population under study.

Interval estimates for the first purpose are called confidence intervals; intervals for the second purpose are called prediction intervals. Though related, the two types of interval estimates are not identical, and cannot be interchanged.

In sections 3.3 and 3.4, confidence intervals will be developed for both the median and mean. Prediction intervals, both parametric and nonparametric, will be used in sections 3.5 and 3.6 to judge whether one new observation is consistent with existing data. Intervals for percentiles other than the median will be discussed in section 3.7.



3.2 Interpretation of Interval Estimates

Suppose that the true population mean μ of concentration in an aquifer was 10. Also suppose that the true population variance σ^2 equals 1. As these values in practice are never known, samples are taken to estimate them by the sample mean $\bar{x}$ and sample variance s^2. Sufficient funding is available to take 12 water samples (roughly one per month) during a year, and the days on which sampling occurs are randomly chosen. From these 12 samples $\bar{x}$ and s^2 are computed. Although in reality only one set of 12 samples would be taken each year, using a computer 12 days can be selected multiple times to illustrate the concept of an interval estimate. For each of 10 independent sets of 12 samples, a confidence interval on the mean is computed using equations given later in section 3.4.1. The results are shown in table 3.1 and figure 3.1.

	N	Mean	St. Dev.	90 % Confidence Interval
1	12	10.06	1.11	(9.49 to 10.64)
2	12	10.60	0.81	*(10.18 to 11.02)
3	12	9.95	1.26	(9.29 to 10.60)
4	12	10.18	1.26	(9.52 to 10.83)
5	12	10.17	1.33	(9.48 to 10.85)
6	12	10.22	1.19	(9.60 to 10.84)
7	12	9.71	1.51	(8.92 to 10.49)
8	12	9.90	1.01	(9.38 to 10.43)
9	12	9.95	0.10	(9.43 to 10.46)
10	12	9.88	1.37	(9.17 to 10.59)

Table 3.1 Ten 90% confidence intervals around a true mean of 10. Data follow a normal distribution. The interval with the asterisk does not include the true value.

These ten intervals are "90% confidence intervals" on the true population mean. That is, the true mean will be contained in these intervals an average of 90 percent of the time. So for the 10 intervals in the table, nine are expected to include the true value while one is not. This is in

fact what happened. Of course when a one-time sampling occurs, the computed interval will either include or not include the true, unknown population mean. The probability that the interval does include the true value is called the **confidence level** of the interval. The probability that this interval will not cover the true value, called the **alpha level** (α), is computed as

$$\alpha = 1 - \text{confidence level.} \qquad [3.1]$$

The width of a confidence interval is a function of the shape of the data distribution (its variability and skewness), the sample size, and of the confidence level desired. As the confidence level increases the interval width also increases, because a larger interval is more likely to contain the true value than is a shorter interval. Thus a 95% confidence interval will be wider than a 90% interval for the same data.

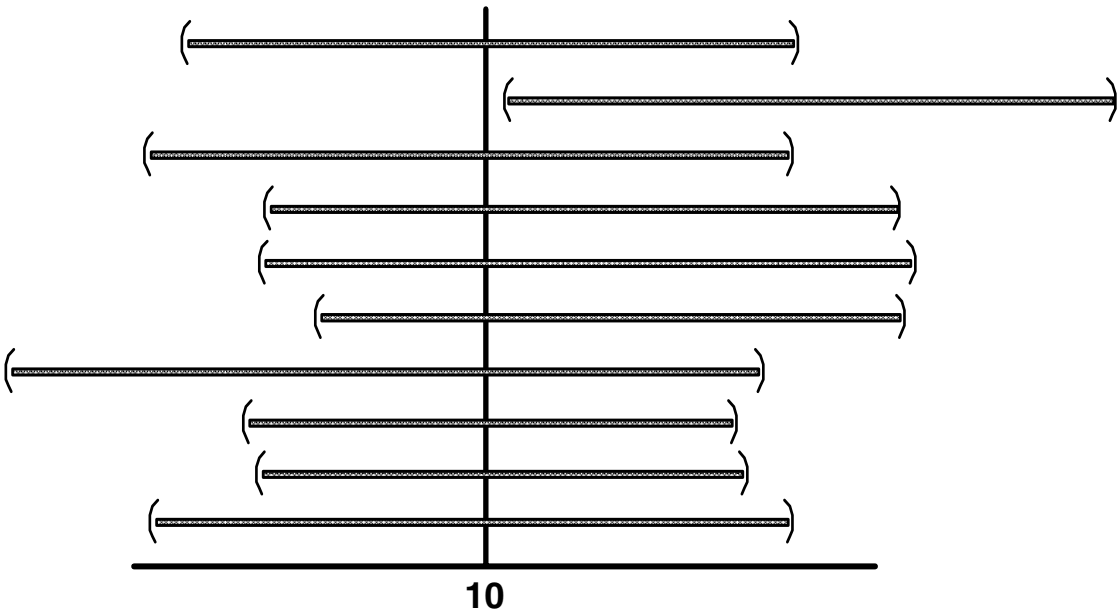


Figure 3.1 Ten 90% confidence intervals for normally-distributed data with true mean = 10

Symmetric confidence intervals on the mean are commonly computed assuming the data follow a normal distribution (see section 3.4.1). If not, the distribution of the mean itself will be approximately normal as long as sample sizes are large (say 50 observations or greater). Confidence intervals assuming normality will then include the true mean $(1-\alpha)$% of the time. In the above example, the data were generated from a normal distribution, so the small sample size of 12 is not a problem. However when data are skewed and sample sizes below 50 or more, symmetric confidence intervals will not contain the mean $(1-\alpha)$% of the time. In the example below, symmetric confidence intervals are incorrectly computed for skewed data (figure 3.2). The results (figure 3.3 and table 3.2) show that the confidence intervals miss the true value of 1 more frequently than they should. The greater the skewness, the larger the sample size must be

before symmetric confidence intervals can be relied on. As an alternative, asymmetric confidence intervals can be computed for the common situation of skewed data. They are also presented in the following sections.

	N	Mean	St. Dev.	90 % Confidence Interval
1	12	0.784	0.320	*(0.618 to 0.950)
2	12	0.811	0.299	*(0.656 to 0.966)
3	12	1.178	0.700	(0.815 to 1.541)
4	12	1.030	0.459	(0.792 to 1.267)
5	12	1.079	0.573	(0.782 to 1.376)
6	12	0.833	0.363	(0.644 to 1.021)
7	12	0.789	0.240	*(0.664 to 0.913)
8	12	1.159	0.815	(0.736 to 1.581)
9	12	0.822	0.365	*(0.633 to 0.992)
10	12	0.837	0.478	(0.589 to 1.085)

Table 3.2 Ten 90% confidence intervals around a true mean of 1. Data do not follow a normal distribution. Intervals with an asterisk do not include the true value.

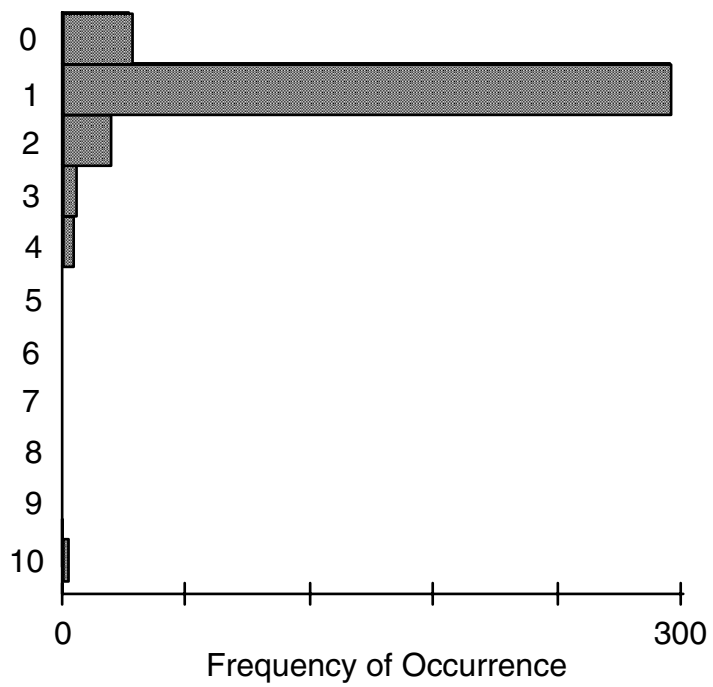


Figure 3.2 Histogram of skewed example data. $\mu = 1.0$ $\sigma = 0.75$.

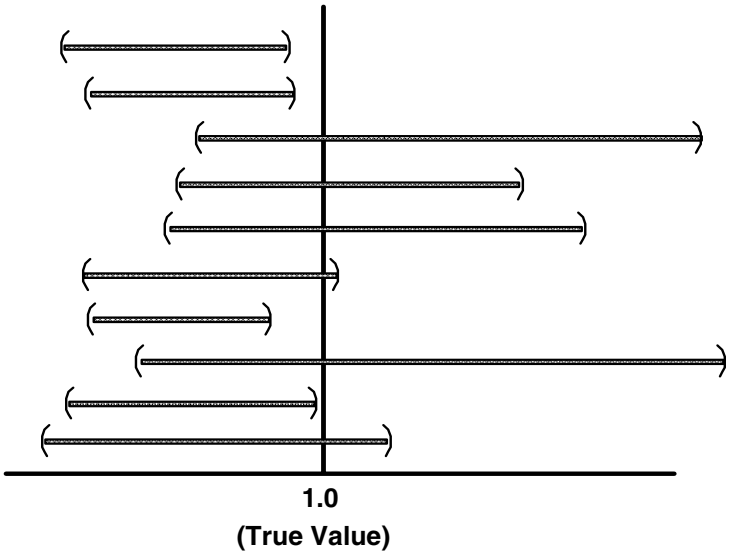
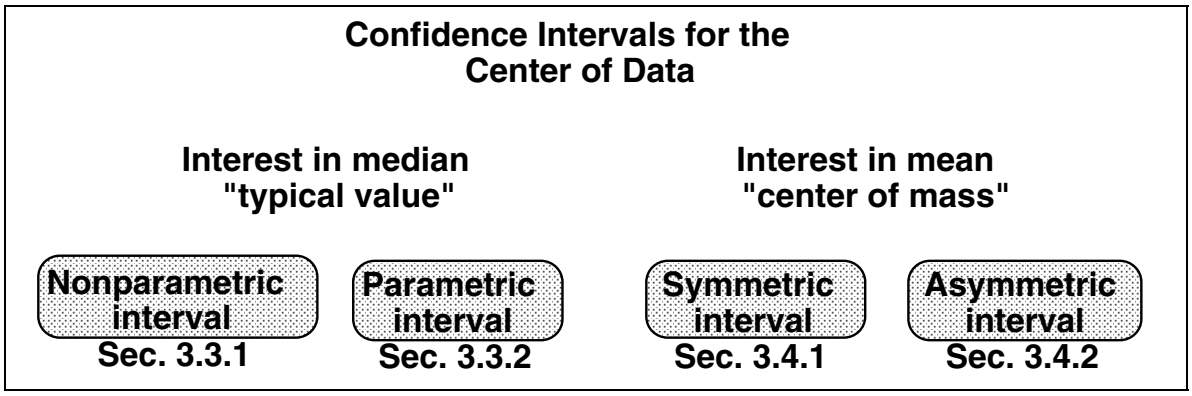


Figure 3.3 Ten 90% confidence intervals for skewed data with true mean = 1.0

3.3 Confidence Intervals for the Median

A confidence interval for the true population median may be computed either without assuming the data follow any specific distribution (section 3.3.1), or assuming they follow a distribution such as the lognormal (section 3.3.2).



3.3.1 Nonparametric Interval Estimate for the Median

A nonparametric interval estimate for the true population median is computed using the binomial distribution. First, the desired significance level α is stated, the acceptable risk of not including the true median. One-half ($\alpha/2$) of this risk is assigned to each end of the interval (figure 3.4). A table of the binomial distribution provides lower and upper critical values x' and x at one-half the desired alpha level ($\alpha/2$). These critical values are transformed into the ranks R_l and R_u corresponding to data points C_l and C_u at the ends of the confidence interval.

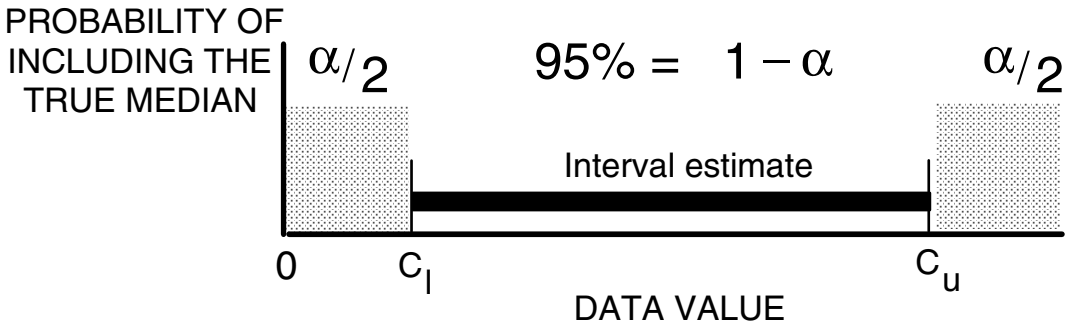


Figure 3.4 Probability of containing the true median $P_{.50}$ in a 2-sided interval estimate. $P_{.50}$ will be below the lower interval bound (C_l) $\alpha/2$% of the time, and above the upper interval bound (C_u) $\alpha/2$% of the time.

For small sample sizes, the binomial table is entered at the p=0.5 (median) column in order to compute a confidence interval on the median. This column is reproduced in Appendix Table B5 -- it is identical to the quantiles for the sign test (see chapter 6). A critical value x' is obtained from Table B5 corresponding to $\alpha/2$, or as close to $\alpha/2$ as possible. This critical value is then used to compute the ranks R_u and R_l corresponding to the data values at the upper and lower confidence limits for the median. These limits are the R_lth ranked data points going in from each end of the sorted list of n observations. The resulting confidence interval will reflect the shape (skewed or symmetric) of the original data.

$$R_l = x' + 1 \qquad [3.2]$$

$$R_u = n - x' = x \qquad \text{for } x' \text{ and } x \text{ from Appendix Table B5} \qquad [3.3]$$

Nonparametric intervals cannot always exactly produce the desired confidence level when sample sizes are small. This is because they are discrete, jumping from one data value to the next at the ends of the intervals. However, confidence levels close to those desired are available for all but the smallest sample sizes.

<u>Example 2</u>

The following 25 arsenic concentrations (in ppb) were reported for ground waters of southeastern New Hampshire (Boudette and others, 1985). A histogram of the data is shown in figure 3.5. Compute the α=0.05 interval estimate of the median concentration.

1.3	1.5	1.8	2.6	2.8	3.5	4.0	4.8	8
9.5	12	14	19	23	41	80	100	110
120	190	240	250	300	340	580		

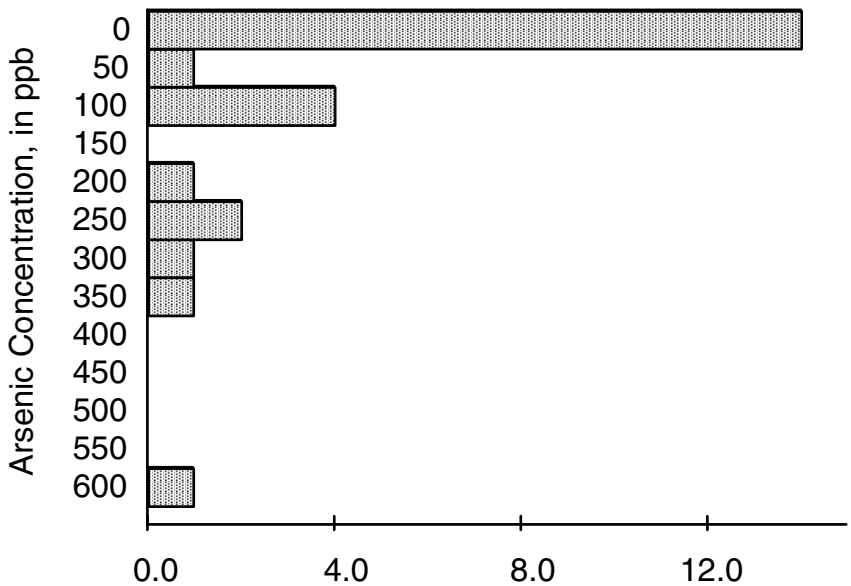


Figure 3.5 Histogram of Example 2 arsenic concentrations (in ppb)

The sample median concentration $\hat{C}_{0.5} = 19$, the 13th observation ranked from smallest to largest. To determine a 95% confidence interval for the true median concentration $C_{0.5}$, the tabled critical value with an entry nearest to $\alpha/2 = 0.025$ is x' = 7 from Table B5. The entry value of 0.022 is quite near 0.025, and is the equivalent to the shaded area at one side of figure 3.4. From equations 3.2 and 3.3 the rank R_l of the observation corresponding to the lower confidence limit is 8, and R_u corresponding to the upper confidence limit is $25 - 7 = 18$.

For this confidence interval the alpha level $\alpha = 2 \bullet 0.022 = 0.044$. This is equivalent to a 1−0.044 or 95.6% confidence limit for $C_{0.5}$, and is the interval between the 8th and 18th ranked observations (the 8th point in from either end), or

$$C_l = 4.8 \leq C_{0.5} \leq 110 = C_u \quad \text{at } \alpha = 0.044$$

The asymmetry around $\hat{C}_{0.5} = 19$ reflects the skewness of the data.

An alternative method for computing the same nonparametric interval is used when the sample size n>20. This large-sample approximation utilizes a table of the standard normal distribution available in every basic statistics textbook to approximate the binomial distribution. By using this approximation, only small tables of the binomial distribution up to n=20 need be included in statistics texts. A critical value $z_{\alpha/2}$ from the normal table determines the upper and lower ranks of observations corresponding to the ends of the confidence interval. Those ranks are

$$R_l = \frac{n - z_{\alpha/2}\sqrt{n}}{2} \qquad [3.4]$$

$$R_u = \frac{n + z_{\alpha/2}\sqrt{n}}{2} + 1 \qquad [3.5]$$

The computed ranks R_u and R_l are rounded to the nearest integer when necessary.

<u>Example 2, cont.</u>
For the n=25 arsenic concentrations, an approximate 95 percent confidence interval on the true median $C_{0.5}$ is computed using $z_{\alpha/2} = 1.96$ so that

$$R_l = \frac{25 - 1.96 \bullet \sqrt{25}}{2} = 7.6$$

$$R_u = \frac{25 + 1.96 \bullet \sqrt{25}}{2} + 1 = 18.4$$

the "7.6th ranked observation" in from either end. Rounding to the nearest integer, the 8th and 18th ranked observations are used as the ends of the α=0.05 confidence limit on $C_{0.5}$, agreeing with the exact 95.6% confidence limit computed previously.

3.3.2 Parametric Interval Estimate for the Median
As mentioned in chapter 1, the geometric mean of x (GM_x) is an estimate of the median in original (x) units when the data logarithms y = ln(x) are symmetric. The mean of y and confidence interval on the mean of y become the geometric mean with its (asymmetric) confidence interval after being retransformed back to original units by exponentiation (equations 3.6 and 3.7). These are parametric alternatives to the point and interval estimates of section 3.3.1. Here it is assumed that the data are distributed as a lognormal distribution. The geometric mean and interval would be more efficient (shorter interval) measures of the median and its confidence interval when the data are truly lognormal. The sample median and its interval are more appropriate and more efficient if the logarithms of data still exhibit skewness and/or outliers.

$$GM_x = \exp(\bar{y}) \qquad \text{where } y = \ln(x) \text{ and } \bar{y} = \text{sample mean of y.} \qquad [3.6]$$

$$\exp\left(\bar{y} - t_{(\alpha/2,n-1)}\sqrt{s_y^2/n}\right) \le GM_x \le \exp\left(\bar{y} - t_{(\alpha/2,n-1)}\sqrt{s_y^2/n}\right) \qquad [3.7]$$

where s_y^2 = sample variance of y in natural log units.

Example 2, cont.

Natural logs of the arsenic data are as follows:

0.262	0.405	0.588	0.956	1.030	1.253	1.387	1.569	2.079
2.251	2.485	2.639	2.944	3.135	3.714	4.382	4.605	4.700
4.787	5.247	5.481	5.521	5.704	5.829	6.363		

The mean of the logs = 3.17, with standard deviation of 1.96. From figure 3.6 the logs of the data appear more symmetric than do the original units of concentration shown previously in figure 3.5.

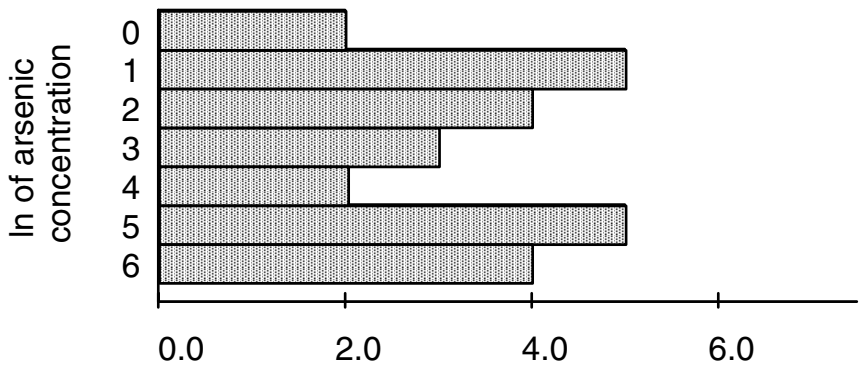


Figure 3.6 Histogram of natural logs of the arsenic concentrations of Example 2

From equations 3.6 and 3.7, the geometric mean and its 95% confidence interval are:

$$GM_C = \exp(3.17) = 23.8$$

$$\exp\left(3.17 - 2.064 \bullet \sqrt{1.96^2/25}\right) \le GM_C \le \exp\left(3.17 + 2.064 \bullet \sqrt{1.96^2/25}\right)$$

$$\exp(2.36) \le GM_C \le \exp(3.98)$$

$$10.6 \le GM_C \le 53.5$$

The scientist must decide whether it is appropriate to assume a lognormal distribution. If not, the nonparametric interval of section 3.3.1 would be preferred.

3.4 Confidence Intervals for the Mean

Interval estimates may also be computed for the true population mean μ. These are appropriate if the center of mass of the data is the statistic of interest (see Chapter 1). Intervals symmetric around the sample mean $\overline{X}$ are computed most often. For large sample sizes a symmetric interval adequately describes the variation of the mean, regardless of the shape of the data distribution. This is because the distribution of the sample mean will be closely approximated by

a normal distribution as sample sizes get larger, even though the data may not be normally distributed†. For smaller sample sizes, however, the mean will not be normally distributed unless the data themselves are normally distributed. As data increase in skewness, more data are required before the distribution of the mean can be adequately approximated by a normal distribution. For highly skewed distributions or data containing outliers, it may take more than 100 observations before the mean will be sufficiently unaffected by the largest values to assume that its distribution will be symmetric.

3.4.1 Symmetric Confidence Interval for the Mean

Symmetric confidence intervals for the mean are computed using a table of the student's t distribution available in statistics textbooks and software. This table is entered to find critical values for t at one-half the desired alpha level. The width of the confidence interval is a function of these critical values, the standard deviation of the data, and the sample size. When data are skewed or contain outliers, the assumptions behind the t-interval do not hold. The resulting symmetric interval will be so wide that most observations will be included in it. It may also extend below zero on the lower end. Negative endpoints of a confidence interval for data which cannot be negative are clear signals that the assumption of a symmetric confidence interval is not warranted. For such data, assuming a lognormal distribution as described in section 3.4.2 would be more appropriate.

The student's t statistic $t_{(\alpha/2,\, n-1)}$ is used to compute the following symmetric confidence interval:

$$\bar{x} - t_{(\alpha/2,\, n-1)} \bullet \sqrt{s^2/n} \;\le\; \mu \;\le\; \bar{x} + t_{(\alpha/2,\, n-1)} \bullet \sqrt{s^2/n} \qquad [3.8]$$

<u>Example 2, cont.</u>

The sample mean arsenic concentration $\bar{C} = 98.4$. This is the point estimate for the true unknown population mean μ. An $\alpha = 0.05$ confidence interval on μ is

$$98.4 - t_{(.025,\, 24)} \bullet \sqrt{144.7^2/25} \;\le \mu \le\; 98.4 + t_{(.025,\, 24)} \bullet \sqrt{144.7^2/25}$$
$$98.4 - 2.064 \bullet 28.9 \;\le \mu \le\; 98.4 + 2.064 \bullet 28.9$$
$$38.7 \;\le \mu \le\; 158.1$$

Thus there is a 95% probability that μ is contained in the interval between 38.7 and 158.1 ppb, assuming that a symmetric confidence interval is appropriate. Note that this confidence interval is, like $\bar{C}$, sensitive to the highest data values. If the largest value of 580 were changed to 380, the median and its confidence interval would be unaffected. $\bar{C}$ would change to 90.4, with a 95% interval estimate for μ from 40.7 to 140.1.

---

† This property is called the Central Limit Theorem (Conover, 1980). It holds for data which follow a distribution having finite variance, and so includes most distributions of interest in water resources.

3.4.2 Asymmetric Confidence Interval for the Mean (for Skewed Data)

Means and confidence intervals may also be computed by assuming that the logarithms y = ln(x) of the data are symmetric. If the data appear more like a lognormal than a normal distribution, this assumption will give a more reliable (lower variance) estimate of the mean than will computation of the usual sample mean without log transformation.

To estimate the population mean μ_X in original units, assume the data are lognormal. One-half the variance of the logarithms is added to $\bar{y}$ (the mean of the logs) prior to exponentiation (Aitchison and Brown, 1981). As the sample variance s^2_y is only an estimate of the true variance of the logarithms, the sample estimate of the mean is biased (Bradu and Mundlak, 1970). However, for small s^2_y and large sample sizes the bias is negligible. See Chapter 9 for more information on the bias of this estimator.

$$\hat{\mu}_x = \exp(\bar{y} + 0.5 \bullet s^2_y) \qquad \text{where } y = \ln(x), \qquad [3.9]$$

$\bar{y}$ = sample mean and s^2_y = sample variance of y in natural log units.

The confidence interval around $\hat{\mu}_x$ is <u>not</u> the interval estimate computed for the geometric mean in equation 3.7. It cannot be computed simply by exponentiating the interval around $\bar{y}$. An exact confidence interval in original units for the mean of lognormal data can be computed, though the equation is beyond the scope of this book. See Land (1971) and (1972) for details.

<u>Example 2, cont.</u>

To estimate the mean concentration assuming a lognormal distribution,

$$\hat{\mu}_c = \exp(3.17 + 0.5 \bullet 1.96^2) = 162.8 .$$

This estimate does not even fall within the confidence interval computed earlier for the geometric mean ($10.6 \leq GM_C \leq 53.5$). Thus here is a case where it is obvious that the CI on the geometric mean is not an interval estimate of the mean. It is an interval estimate of the median, assuming the data follow a lognormal distribution.

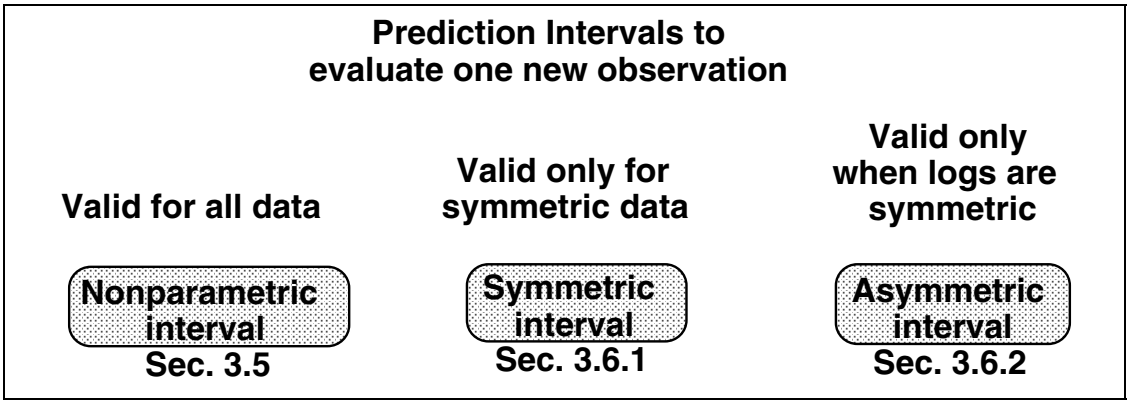
3.5. Nonparametric Prediction Intervals

The question is often asked whether one new observation is likely to have come from the same distribution as previously-collected data, or alternatively from a different distribution. This can be evaluated by determining whether the new observation is outside the **prediction interval** computed from existing data. Prediction intervals contain $100 \bullet (1-\alpha)$ percent of the data distribution, while $100 \bullet \alpha$ percent are outside of the interval. If a new observation comes from the same distribution as previously-measured data, there is a $100 \bullet \alpha$ percent chance that it will lie

outside of the prediction interval. Therefore being outside the interval does not "prove" the new observation is different, just that it is likely to be so. How likely this is depends on the choice of α made by the scientist.

Prediction intervals are computed for a different purpose than confidence intervals -- they deal with individual data values as opposed to a summary statistic such as the mean. A prediction interval is wider than the corresponding confidence interval, because an individual observation is more variable than is a summary statistic computed from several observations. Unlike a confidence interval, a prediction interval takes into account the variability of single data points around the median or mean, in addition to the error in estimating the center of the distribution. When the mean $\pm$ 2 standard deviations are mistakenly used to estimate the width of a prediction interval, new data are asserted as being from a different population more frequently than they should.

In this section nonparametric prediction intervals are presented -- intervals not requiring the data to follow any particular distributional shape. Prediction intervals can also be developed assuming the data follow a particular distribution, such as the normal. These are discussed in section 3.6. Both two-sided and one-sided prediction intervals are described.



It may also be of interest to know whether the median or mean of a new set of data differs from that for an existing group. To test for differences in medians, use the rank-sum test of Chapter 5. To test for differences in means, the two-sample t-test of Chapter 5 should be performed.

3.5.1 Two-Sided Nonparametric Prediction Interval

The nonparametric prediction interval of confidence level α is simply the interval between the $\alpha/2$ and $1-(\alpha/2)$ percentiles of the distribution (figure 3.7). This interval contains $100\bullet(1-\alpha)$ percent of the data, while $100\bullet\alpha$ percent lies outside of the interval. Therefore if the new additional data point comes from the same distribution as the previously measured data, there is a $100\bullet\alpha$ percent chance that it will lie outside of the prediction interval and be incorrectly

labeled as "changed". The interval will reflect the shape of the data it is developed from, and no assumptions about the form of that shape need be made.

$$PI_{np} = X_{\alpha/2 \bullet (n+1)} \quad \text{to} \quad X_{[1-(\alpha/2)] \bullet (n+1)} \qquad [3.10]$$

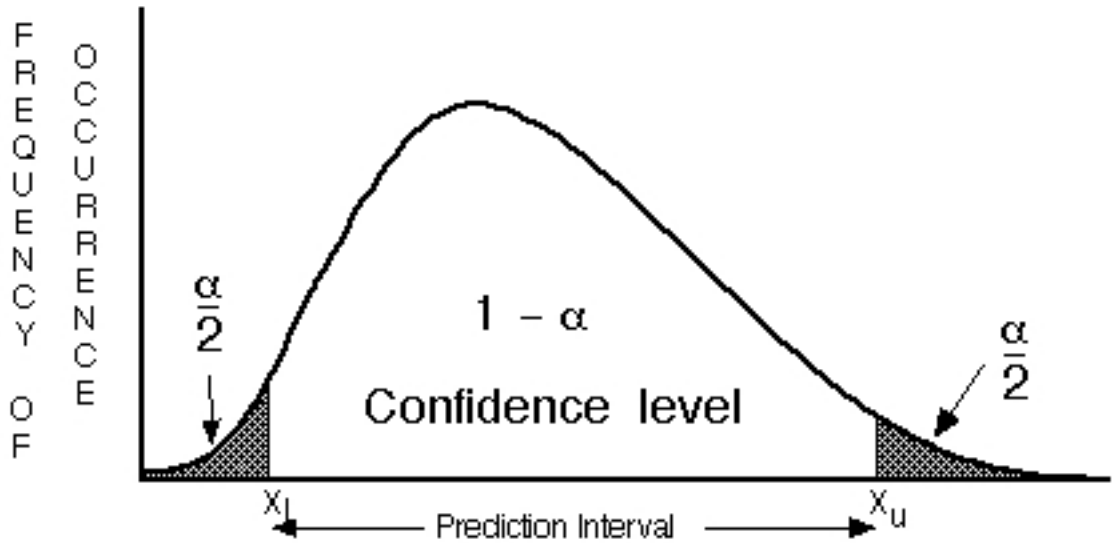


Figure 3.7 Two-sided prediction interval. A new observation will be below X_l $\alpha/2$% and above X_u $\alpha/2$% of the time, when the data distribution is unchanged.

<u>Example 2, cont.</u>
Compute a 90% (α = 0.10) prediction interval for the arsenic data without assuming the data follow any particular distribution.

The 5th and 95th percentiles of the arsenic data are the observations with ranks of (.05•26) and (.95•26), or 1.3 and 24.7. By linearly interpolating between the 1st and 2nd, and 24th and 25th observations, the α = 0.10 prediction interval is

$$X_1 + 0.3 \bullet (X_2 - X_1) \quad \text{to} \quad X_{24} + 0.7 \bullet (X_{25} - X_{24})$$
$$1.3 + 0.3 \bullet 0.2 \quad \text{to} \quad 340 + 0.7 \bullet 240$$
$$1.4 \quad \text{to} \quad 508 \text{ ppb}$$

A new observation less than 1.4 or greater than 508 can be considered as coming from a different distribution at a 10% risk level (α = 0.10).

3.5.2 One-Sided Nonparametric Prediction Interval

One-sided prediction intervals are appropriate if the interest is in whether a new observation is larger than existing data, or smaller than existing data, but not both. The decision to use a one-sided interval must be based entirely on the question of interest. It should not be determined

after looking at the data and deciding that the new observation is likely to be only larger, or only smaller, than existing information. One-sided intervals use α rather than $\alpha/2$ as the error risk, placing all the risk on one side of the interval (figure 3.8).

one-sided PI $_{np}$:	new x < $X_{\alpha\bullet(n+1)}$, or new x > $X_{[1-\alpha]\bullet(n+1)}$ (but not either, or)	[3.11]

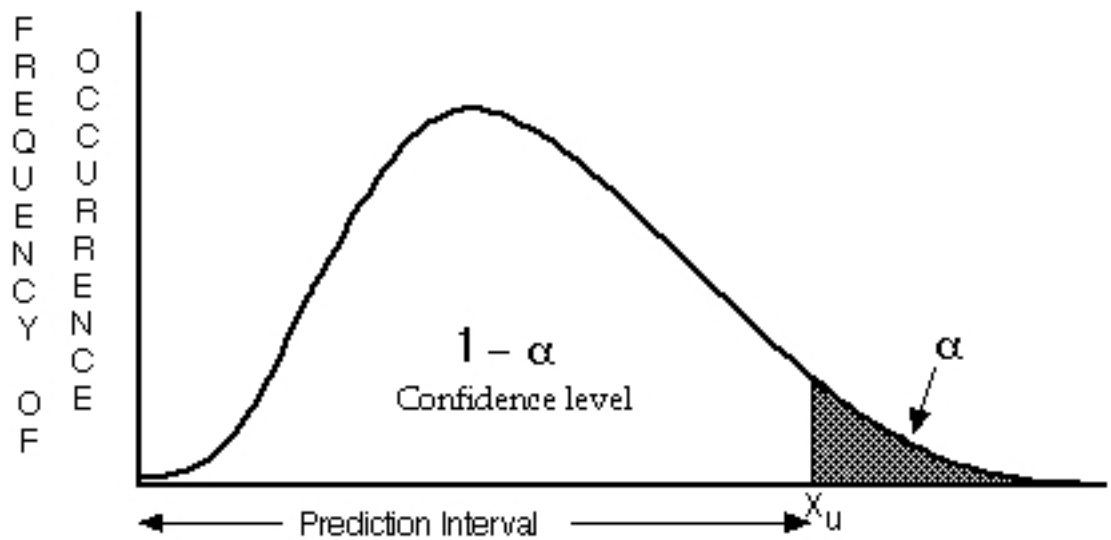


Figure 3.8 Confidence level and alpha level for a 1-sided prediction interval Probability of obtaining a new observation greater than X_u when the distribution is unchanged is α.

<u>Example 2, cont.</u>
An arsenic concentration of 350 ppb is found in a New Hampshire well. Does this indicate a change to larger values as compared to the distribution of concentrations for the example 2 data? Use $\alpha = 0.10$.

As only large concentrations are of interest, the new data point will be considered larger if it exceeds the $\alpha = 0.10$ one-sided prediction interval, or upper 90th percentile of the existing data. $X_{0.90\bullet 26} = X_{23.4}$. By linear interpolation this corresponds to a concentration of

$$X_{23} + 0.4\bullet(X_{24}-X_{23}) \;=\; 300 + 0.4\bullet(40) \;=\; 316.$$

In other words, a concentration of 316 or greater will occur approximately 10 percent of the time if the distribution of data has <u>not</u> increased. Therefore a concentration of 350 ppb is considered larger than the existing data at an α level of 0.10.

3.6 Parametric Prediction Intervals

Parametric prediction intervals are also used to determine whether a new observation is likely to come from a different distribution than previously-collected data. However, an assumption is now made about the shape of that distribution. This assumption provides more information with which to construct the interval, as long as the assumption is valid. If the data do not approximately follow the assumed distribution, the prediction interval may be quite inaccurate.

3.6.1 Symmetric Prediction Interval

The most common assumption is that the data follow a normal distribution. Prediction intervals are then constructed to be symmetric around the sample mean, and wider than confidence intervals on the mean. The equation for this interval differs from that for a confidence interval around the mean by adding a term $\sqrt{s^2}$ = s, the standard deviation of individual observations around their mean:

$$\text{PI} = \bar{X} - t_{(\alpha/2,\, n-1)} \bullet \sqrt{s^2 + (s^2/n)} \quad \text{to} \quad \bar{X} + t_{(\alpha/2,\, n-1)} \bullet \sqrt{s^2 + (s^2/n)} \qquad [3.12]$$

One-sided intervals are computed as before, using α rather than $\alpha/2$ and comparing new data to only one end of the prediction interval.

<u>Example 2, cont.</u>

Assuming symmetry, is a concentration of 350 ppb different (not just larger) than what would be expected from the previous distribution of arsenic concentrations? Use $\alpha = 0.10$.

The parametric two-sided $\alpha = 0.10$ prediction interval is

$$98.4 - t_{(.05,\, 24)} \bullet \sqrt{144.7^2 + 144.7^2/25} \quad \text{to} \quad 98.4 + t_{(.05,\, 24)} \bullet \sqrt{144.7^2 + 144.7^2/25}$$

$$98.4 - 1.711 \bullet 147.6 \quad \text{to} \quad 98.4 + 1.711 \bullet 147.6$$

$$-154.1 \quad \text{to} \quad 350.9$$

350 ppb is at the upper limit of 350.9, so the concentration is not declared different at $\alpha = 0.10$. The negative concentration reported as the lower prediction bound is a clear indication that the underlying data are not symmetric, as concentrations are non-negative. To avoid an endpoint as unrealistic as this negative concentration, an asymmetric prediction interval should be used instead.

3.6.2 Asymmetric Prediction Intervals

Asymmetric intervals can be computed either using the nonparametric intervals of section 3.5, or by assuming symmetry of the logarithms and computing a parametric interval on the logs of the data. Either asymmetric interval is more valid than a symmetric interval when the underlying data are not symmetric, as is the case for the arsenic data of example 2. As stated in Chapter 1,

most water resources data and indeed most environmental data show positive skewness. Thus they should be modelled using asymmetric intervals. Symmetric prediction intervals should be used only when the data are known to come from a normal distribution. This is because prediction intervals deal with the behavior of individual observations. Therefore the Central Limit Theorem (see first footnote in this chapter) does not apply. Data must be assumed non-normal unless shown otherwise. It is difficult to disprove normality using hypothesis tests (Chapter 4) due to the small sample sizes common to environmental data sets. It is also difficult to see non-normality with graphs unless the departures are strong (Chapter 10). It is unfortunate that though most water resources data sets are asymmetric and small, symmetric intervals are commonly used.

An asymmetric (but parametric) prediction interval can be computed using logarithms. This interval is parametric because percentiles are computed assuming that the data follow a lognormal distribution. Thus from equation 3.12:

$$\text{PI} = \exp\left(\bar{y} - t_{(a/2,n-1)}\sqrt{s_y^2 + s_y^2/n}\right) \text{ to } \exp\left(\bar{y} + t_{(a/2,n-1)}\sqrt{s_y^2 + s_y^2/n}\right)$$

where y = ln(X), $\bar{y}$ is the mean and s_y^2 the variance of the logarithms. [3.13]

Example 2, cont.

An asymmetric prediction interval is computed using the logs of the arsenic data. A 90% prediction interval becomes

$$\ln(\text{PI}):\ 3.17 - t_{(0.05,\,24)} \bullet \sqrt{1.96^2 + 1.96^2/25} \quad \text{to} \quad 3.17 + t_{(0.05,\,24)} \bullet \sqrt{1.96^2 + 1.96^2/25}$$

$$3.17 - 1.71 \bullet 2.11 \quad \text{to} \quad 3.17 + 1.71 \bullet 2.11$$

$$0.44 \text{ to } 6.78$$

which when exponentiated into original units becomes

$$1.55 \quad \text{to} \quad 880.1$$

As percentiles can be transformed directly from one measurement scale to another, the prediction interval in log units can be directly exponentiated to give the prediction interval in original units. This parametric prediction interval differs from the one based on sample percentiles in that a lognormal distribution is assumed. The parametric interval would be preferred if the assumption of a lognormal distribution is believed. The sample percentile interval would be preferred when a robust interval is desired, such as when a lognormal model is not believed, or when the scientist does not wish to assume any model for the data distribution.

3.7 Confidence Intervals for Percentiles (Tolerance Intervals)

Quantiles or percentiles have had the traditional use in water resources of describing the frequency of flood events. Thus the 100-year flood is the 99th percentile (0.99 quantile) of the distribution of annual flood peaks. It is the magnitude of flood which is expected to be exceeded only once in 100 years. The 20-year flood is of a magnitude which is expected to be exceeded only once in 20 years (5 times in 100 years), or is the 95th percentile of annual peaks. Similarly, the 2-year flood is the median or 50th percentile of annual peaks. Flood percentiles are determined assuming that peak flows follow a specified distribution. The log Pearson Type III is often used in the United States. Historically, European countries have used the Gumbel (extreme value) distribution, though the GEV distribution is now more common (Ponce, 1989).

The most commonly-reported statistic for analyses of low flows is also based on percentiles, the "7-day 10-year low flow" or 7Q10. The 7Q10 is the 10th percentile of the distribution of annual values of Y, where Y is the lowest average of mean daily flows over any consecutive 7-day period for that year. Y values are commonly fit to Log Pearson III or Gumbel distributions in order to compute the percentile. Often a series of duration periods is used to better define flow characteristics, ie. the 30Q10, 60Q10, and others (Ponce, 1989).

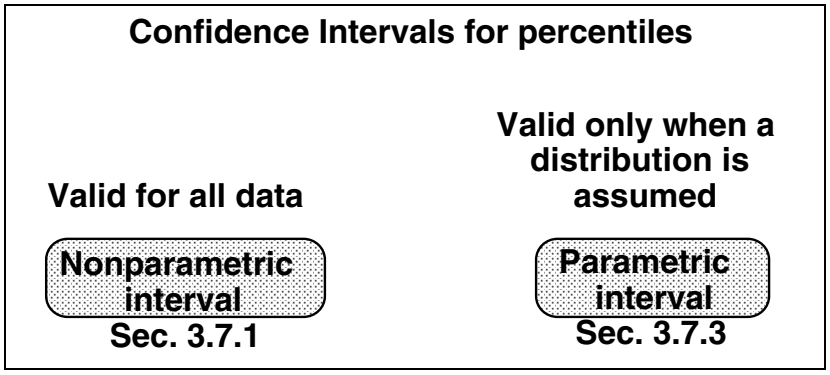
Recently, percentiles: water quality of water-quality records appear to be becoming more important in a regulatory framework. Crabtree et al. (1987) among others have reported an increasing reliance on percentiles for developing and monitoring compliance with water quality standards†. In these scenarios, the median, 95th, or some other percentile should not exceed (or be below) a standard. As of now, no distribution is usually assumed for water-quality concentrations, so that sample percentiles are commonly computed and compared to the standard. In regulatory frameworks, exceedance of a tolerance interval on concentration is sometimes used as evidence of contamination. A tolerance interval is nothing other than a confidence interval on the percentile. The percentile used is the 'coverage coefficient' of the tolerance interval.

In light of the ever increasing use of percentiles in water resources applications, understanding of their variability is quite important. In 3.7.1, interval estimates will be computed without assuming a distribution for the data. Estimates of peak flow percentiles computed in this way will therefore differ somewhat in comparison to those computed using a Log Pearson III or Gumbel assumption. Computation of percentile interval estimates when assuming a specific

---

† Data presented by Crabtree et al. (1987) shows that for each of their cases, percentiles of flow and water-quality constituents are best estimated by (nonparametric) sample percentiles rather than by assuming some distribution. However they come to a different conclusion for two constituents (see their Table 2) by assuming that a parametric process is better unless proven otherwise. In those two cases either could be used.

distributional shape is discussed in section 3.7.3. In sections 3.7.2 and 3.7.4, use of interval estimates for testing hypotheses is illustrated.



3.7.1 Nonparametric Confidence Intervals for Percentiles

Confidence intervals can be developed for any percentile analogous to those developed in section 3.3 for the median. First the desired confidence level is stated. For small sample sizes a table of the binomial distribution is entered to find upper and lower limits corresponding to critical values at one-half the desired alpha level ($\alpha/2$). These critical values are transformed into the ranks corresponding to data points at the ends of the confidence interval.

The binomial table is entered at the column for p, the percentile (actually the quantile) for which a confidence interval is desired. So for a confidence interval on the 75th percentile, the p=0.75 column is used. Go down the column until the appropriate sample size n is found. The tabled probability p* should be found which is as close to $\alpha/2$ as possible. The lower critical value x_l is the integer corresponding to this probability p*. A second critical value x_u is similarly obtained by continuing down the column to find a tabled probability $p' \cong (1-\alpha/2)$. These critical values x_l and x_u are used to compute the ranks R_l and R_u corresponding to the data values at the upper and lower ends of the confidence limit (equations 3.14 and 3.15). The resulting confidence level of the interval will equal $(p'-p^*)$.

$$R_l = x_l + 1 \qquad [3.14]$$

$$R_u = x_u \qquad [3.15]$$

<u>Example 2, cont.</u>
For the arsenic concentrations of Boudette and others (1985), determine a 95% confidence interval on $C_{0.20}$, the 20th percentile of concentration (p=0.2).

The sample 20th percentile $\hat{C}_{0.20}$ = 2.9 ppb, the 0.20•(26) = 5.2 smallest observation, or two-tenths of the distance between the 5th and 6th smallest observations. To determine a 95%

confidence interval for the true 20th percentile $C_{0.20}$, the binomial table from a statistics text such as Bhattacharyya and Johnson (1977) is entered at the p = 0.20 column. The integer x_l having an entry nearest to $\alpha/2 = 0.025$ is found to be 1 ($p^* = 0.027$, the error probability for the lower side of the distribution). From equation 3.14 the rank $R_l = 2$. Going further down the column, $p' = 0.983$ for an $x_u = R_u = 9$. Therefore a 95.6% confidence interval (0.983–0.027 = 0.956) for the 20th percentile is the range between the 2nd and 9th observations, or

$$1.5 \le C_{0.20} \le 8 \quad \text{at } \alpha = 0.044$$

The asymmetry around $\hat{C}_{0.20} = 2.9$ reflects the skewness of the data.

When n>20, a large-sample (normal) approximation to the binomial distribution can be used to obtain interval estimates for percentiles. From a table of quantiles of the standard normal distribution, $z_{\alpha/2}$ and $z_{[1-\alpha/2]}$ (the $\alpha/2$th and $[1-\alpha/2]$th normal quantiles) determine the upper and lower ranks of observations corresponding to the ends of the confidence interval. Those ranks are

$$R_l = np + z_{\alpha/2} \bullet \sqrt{np\,(1-p)} + 0.5 \qquad [3.16]$$

$$R_u = np + z_{[1-\alpha/2]} \bullet \sqrt{np\,(1-p)} + 0.5 \qquad [3.17]$$

The 0.5 terms added to each reflect a continuity correction (see Chapter 4) of 0.5 for the lower bound and –0.5 for the upper bound, plus the +1 term for the upper bound analogous to equation 3.5. The computed ranks R_u and R_l are rounded to the nearest integer.

<u>Example 2, cont.</u>

Using the large sample approximation of equations 3.16 and 3.17, what is a 95% confidence interval estimate for the true 0.2 quantile?

Using $z_{\alpha/2} = -1.96$, the lower and upper ranks of the interval are

$$R_l = 25\bullet 0.2 + (-1.96) \bullet \sqrt{25\bullet 0.2\,(1-0.2)} + 0.5 = 5 - 1.96\bullet 2 + 0.5 = 1.6$$

$$R_u = 25\bullet 0.2 + 1.96\bullet \sqrt{25\bullet 0.2\,(1-0.2)} + 0.5 = 5 + 1.96\bullet 2 + 0.5 = 9.4$$

After rounding, the 2nd and 9th ranked observations are found to be an approximate α=0.05 confidence limit on $C_{0.2}$, agreeing with the exact confidence limit computed previously.

3.7.2 Nonparametric Tests for Percentiles

Often it is of interest to test whether a percentile is different from, or larger or smaller than, some specified value. For example, a water quality standard X_0 could be set such that the median of daily concentrations should not exceed X_0 ppb. Or the 10-year flood (90th percentile of annual peak flows) may be tested to determine if it differs from a regional design value X_0. Detailed discussions of hypothesis tests do not begin until the next chapter. However, a simple way to view such a test is discussed below. It is directly related to confidence intervals.

3.7.2.1 N-P test for whether a percentile differs from X_0 (a two-sided test)
To test whether the percentile of a data set is significantly different (either larger or smaller) from a pre-specified value X_0, compute an interval estimate for the percentile as described in section 3.7.1. If X_0 falls within this interval, the percentile is not significantly different from X_0 at a significance level = α (figure 3.9). If X_0 is not within the interval, the percentile significantly differs from X_0 at the significance level of α.

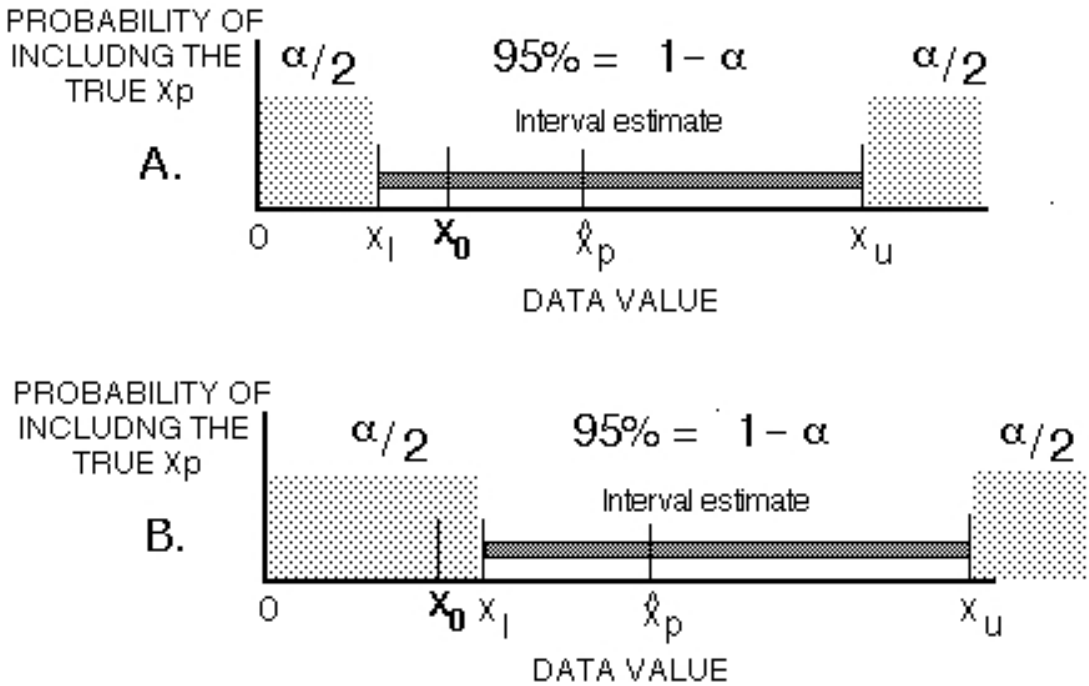


Figure 3.9 Interval estimate of pth percentile X_p as a test for whether $X_p = X_0$.
A. X_0 inside interval estimate: X_p not significantly different from X_0.
B. X_0 outside interval estimate: X_p significantly different from X_0.

<u>Example 3</u>
In Appendix C1 are annual peak discharges for the Saddle R. at Lodi, NJ from 1925 to 1967. Of interest is the 5-year flood, the flood which is likely to be equalled or exceeded once every 5 years (20 times in 100 years), and so is the 80th percentile of annual peaks. Though flood percentiles are usually computed assuming a Log Pearson Type III or Gumbel distribution (Ponce, 1989), here they will be estimated by the sample 80th percentile. Is there evidence that the 20-year flood between 1925-1967 differs from a design value of 1300 cfs at an α= 0.05?

The 80th percentile is estimated from the 43 values between 1925 and 1967 as the 0.8•(44) = 35.2 value when ranked from smallest to largest. Therefore $\hat{Q}_{0.8}$ = 1672 cfs, 0.2

of the distance between the 35th and 36th ranked peak flow. A two-sided confidence interval on this percentile is (following equations 3.16 and 3.17):

$$R_l = np + z_{\alpha/2} \bullet \sqrt{np\,(1\text{-}p)} \;+ 0.5 \qquad\qquad R_u = np + z_{[1-\alpha/2]} \bullet \sqrt{np\,(1\text{-}p)} \;+ 0.5$$

$$R_l = 43(0.8) - 1.96 \bullet \sqrt{43 \bullet 0.8\,(0.2)} \;+ 0.5 = 29.8 \qquad\qquad R_u = 43(0.8) + 1.96 \bullet \sqrt{43 \bullet 0.8\,(0.2)} \;+ 0.5 = 40.0$$

The α=0.05 confidence interval lies between the 30th and 40th ranked peak flows, or

$$1370 < Q_{0.8} < 1860$$

which does not include the design value X_0 = 1300 cfs. Therefore the 20-year flood does differ from the design value at a significance level of 0.05.

3.7.2.2 N-P test for whether a percentile exceeds X_0 (a one-sided test)

To test whether a percentile X_p significantly exceeds a specified value or standard X_0, compute the one-sided confidence interval of section 3.7.1. Remember that the entire error level α is placed on the side below the percentile point estimate $\hat{X}_p$ (figure 3.10). X_p will be declared significantly higher than X_0 if its one-sided confidence interval lies entirely above X_0.

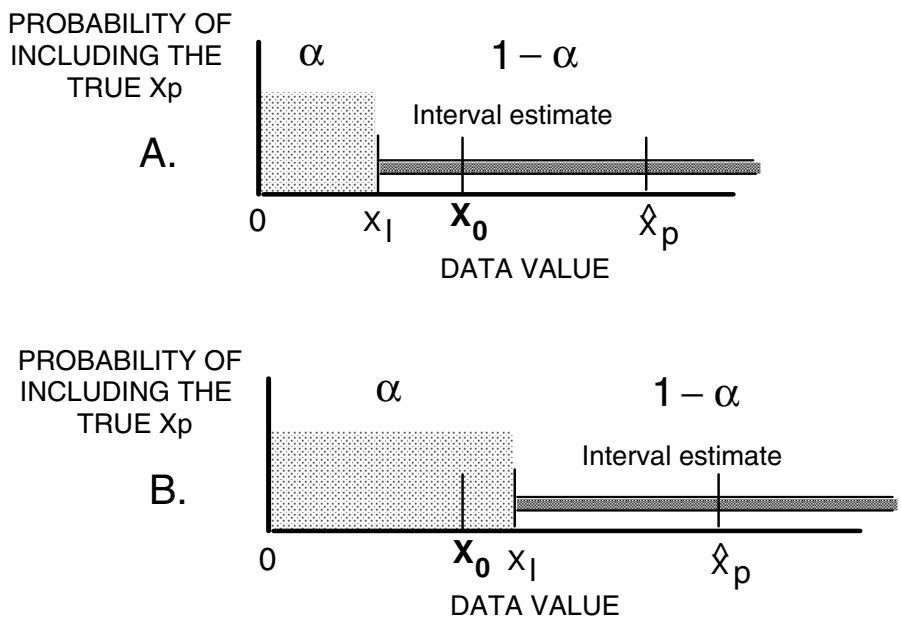


Figure 3.10 One-sided interval estimate as a test for whether percentile $X_p > X_0$.
A. X_0 inside interval estimate: X_p not significantly greater than X_0.
B. X_0 below interval estimate: X_p significantly greater than X_0.

<u>Example 2, cont.</u>

Suppose that a water-quality standard stated that the 90th percentile of arsenic concentrations in drinking water shall not exceed 300 ppb. Has this standard been violated at the $\alpha = 0.05$ confidence level by the New Hampshire data of example 2?

The 90th percentile of the example 2 arsenic concentrations is

$$\hat{C}_{.90} = (25+1)\bullet 0.9\text{th} = 23.4\text{th data point} = 300 + 0.4\,(340-300)$$
$$= 316 \text{ ppb.}$$

Following equation 3.16 but using α instead of $\alpha/2$, the rank of the observation corresponding to a one-sided 95% lower confidence bound on $C_{.90}$ is

$$R_l = np + z_\alpha \bullet \sqrt{np\,(1\text{-}p)} + 0.5 = 25\bullet 0.9 + z_{0.05} \bullet \sqrt{25\bullet 0.9\,(0.1)} + 0.5$$
$$= 22.5 + (-1.64)\bullet\sqrt{2.25} + 0.5$$
$$= 20.5$$

and thus the lower confidence limit is the 20.5th lowest observation, or 215 ppb, halfway between the 20th and 21st observations. This confidence limit is less than X_0 =300, and therefore the standard has not been exceeded at the 95% confidence level.

3.7.2.3 N-P test for whether a percentile is less than X_0 (a one-sided test)

To test whether a percentile X_p is significantly less than X_0, compute the one-sided confidence interval placing all error α on the side above $\hat{X}_p$ (figure 3.11). X_p will be declared as significantly less than X_0 if its one-sided confidence interval is entirely below X_0.

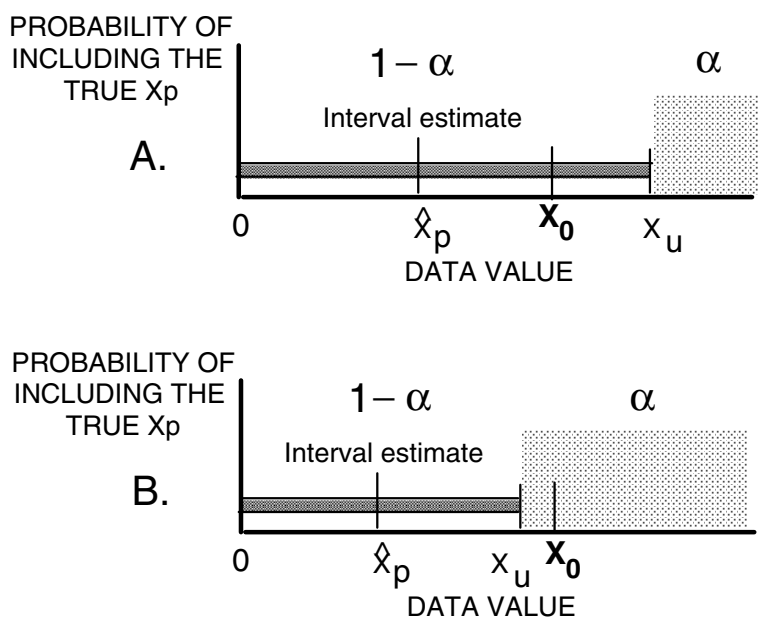


Figure 3.11 One-sided interval estimate as a test for whether percentile $X_p < X_0$.
A. X_0 inside interval estimate: X_p not significantly less than X_0.
B. X_0 above interval estimate: X_p significantly less than X_0.

Example 4

The following 43 values are annual 7-day minimum flows for 1941–1983 on the Little Mahoning Creek at McCormick, PA. Though percentiles of low flows are often computed using a Log Pearson Type III distribution, here the sample estimate of the percentile will be computed. Is the 7Q10 low-flow (the 10th percentile of these data) significantly less than 3 cfs at $\alpha = 0.05$?

0.69	0.80	1.30	1.40	1.50	1.50	1.80	1.80	2.10	2.50	2.80
2.90	3.00	3.10	3.30	3.70	3.80	3.80	4.00	4.10	4.20	4.30
4.40	4.80	4.90	5.70	5.80	5.90	6.00	6.10	7.90	8.00	8.00
9.70	9.80	10.00	11.00	11.00	12.00	13.00	16.00	20.00	23.00	

The sample 10th percentile of the data is 4.4th observation, or $7\hat{Q}_{.10} = 1.4$ cfs. The upper 95% confidence interval for $Q_{.10}$ is located (following equation 3.17 but using α) at rank R_u:

$$\begin{aligned} R_u &= np + z_{[1-\alpha]} \bullet \sqrt{np\,(1\text{-}p)} + 0.5 \\ &= 43 \bullet 0.1 + 1.64 \bullet \sqrt{43 \bullet 0.1\,(0.9)} + 0.5 \\ &= 8.0 \end{aligned}$$

So the upper 95% confidence limit equals 1.8 cfs. This is below the X_0 of 3 cfs, and therefore the 7Q10 is significantly less than 3 cfs at an $\alpha = 0.05$.

3.7.3 Parametric Confidence Intervals for Percentiles

Confidence intervals for percentiles can also be computed by assuming that data follow a particular distribution. Distributional assumptions are employed because there are often insufficient data to compute percentiles with the required precision. Adding information contained in the distribution will increase the precision of the estimate as long as the distributional assumption is a reasonable one. However when the distribution which is assumed does not fit the data well, the resulting estimates are less accurate, and more misleading, than if nothing were assumed. Unfortunately, the situation in which an assumption is most needed, that of small sample sizes, is the same situation where it is difficult to determine whether the data follow the assumed distribution.

There is little theoretical reason why data should follow one distribution over another. As stated in Chapter 1, most environmental data have a lower bound at zero and may have quite large observations differing from the bulk of the data. Distributions fit to such data must posses skewness, such as the lognormal. But few "first principles" can be drawn on to favor one skewed distribution over another. Empirical studies have found that for specific locations and variables certain distributions seem to fit well, and those have become traditionally used. Thus the lognormal, Pearson Type III and Gumbel distributions are commonly assumed in water resources applications.

Computation of point and interval estimates for percentiles assuming a lognormal distribution are straightforward. First the sample mean $\bar{y}$ and sample standard deviation s_y of the logarithms are computed. The point estimate is then

$$\hat{X}_p = \exp(\bar{y} + z_p \bullet s_y) \qquad [3.18]$$

where z_p is the pth quantile of the standard normal distribution and y = ln(x).

The interval estimate for the median was previously given by equation 3.7 assuming that data are lognormal. For other percentiles, confidence intervals are computed using the non-central t-distribution (Stedinger, 1983). Tables of that distribution are found in Stedinger's article, with more complete entries online in commercial computer mathematical libraries. The confidence interval on X_p is:

$$CI(X_p) = \exp(\bar{y} + \zeta_{\alpha/2} \bullet s_y,\ \bar{y} + \zeta_{[1-\alpha/2]} \bullet s_y) \qquad [3.19]$$

where $\zeta_{\alpha/2}$ is the $\alpha/2$ quantile of the non-central t distribution for the desired percentile with sample size of n.

<u>Example 2, cont.</u>
Compute a 90% interval estimate for the 90th percentile of the New Hampshire arsenic concentrations, assuming the data are lognormal.

The 90th percentile assuming concentrations are lognormal is as given in equation 3.18:

$$\hat{C}_{.90} = \exp(\bar{y} + z_{.90} \bullet s_y) \qquad = \exp(3.17 + 1.28 \bullet 1.96)$$
$$= 292.6 \text{ ppb.}$$

(which is lower than the sample estimate of 316 ppb obtained without assuming the data are lognormal).

The corresponding 90% interval estimate from equation 3.19 is:

$$\exp(\bar{y} + \zeta_{0.05} \bullet s_y) < C_{.90} < \exp(\bar{y} + \zeta_{0.95} \bullet s_y)$$
$$\exp(3.17 + 0.898 \bullet 1.96) < C_{.90} < \exp(3.17 + 1.838 \bullet 1.96)$$
$$138.4 < C_{.90} < 873.5$$

This estimate would be preferred over the nonparametric estimate if it was believed that the data were truly lognormal. Otherwise a nonparametric interval would be preferred. When the data are truly lognormal, the two intervals should be quite similar.

Interval estimates for percentiles of the Log Pearson III distribution are computed in a similar fashion. See Stedinger (1983) for details on the procedure.

3.7.4 Parametric Tests for Percentiles

Analogous to section 3.7.2, parametric interval estimates may be used to conduct a parametric test for whether a percentile is different from (2-sided test), exceeds (1-sided test), or is less than (1-sided test) some specified value X_0. With the 2-sided test for difference, if X_0 falls within the interval having $\alpha/2$ on either side, the percentile is not proven to be significantly different from X_0. If X_0 falls outside this interval, the evidence supports $X_p \neq X_0$ with an error level of α. For the one-sided tests, the error level α is placed entirely on one side before conducting the test, and X_0 is again compared to the end of the interval to determine difference or similarity.

Example 2, cont.

Test whether the 90th percentile of arsenic concentrations in drinking water exceeds 300 ppb at the $\alpha = 0.05$ significance level, assuming the data are lognormal.

The one-sided 95% lower confidence limit for the 90th percentile was computed above as 138.4 ppb. (note the nonparametric bound was 215 ppb). This limit is less than the p_0 value of 300, and therefore the standard has not been exceeded at the 95% confidence level.

3.8 Other Uses for Confidence Intervals

Confidence intervals are used for purposes other than as interval estimates. Three common uses are to detect outliers, for quality control charts, and for determining sample sizes necessary to achieve a stated level of precision. Often overlooked are the implications of data non-normality for the three applications. These are discussed in the following three sections.

3.8.1 Implications of Non-Normality for Detection of Outliers

An outlier is an observation which appears to differ in its characteristics from the bulk of the data set to which it is assigned. It is a subjective concept -- different people may define specific points as either outliers, or not. Outliers are sometimes deleted from a data set in order to use procedures based on the normal distribution. One of the central themes of this book is that this is a dangerous and unwarranted practice. It is dangerous because these data may well be totally valid. There is no law stating that observed data must follow some specific distribution, such as the normal. Outlying observations are often the most important data collected, providing insight into extreme conditions or important causative relationships. Deleting outliers is unwarranted because procedures not requiring an assumption of normality are both available and powerful. Many of these are discussed in the following chapters.

In order to delete an outlier, an observation must first be declared to be one. Rules or "tests" for outliers have been used for years, as surveyed by Beckman and Cook (1983). The most common tests are based on a t-interval, and assume that data follow a normal distribution.

Usually equation 3.12 for a normal prediction interval is simplified by assuming the (s^2/n) terms under the square root sign are negligable compared to s^2 (true for large n). Points beyond the simplified prediction interval are declared as outliers, and dropped.

Real world data may not follow a normal distribution. As opposed to a mean of large data sets, there is no reason to assume that they should. Rejection of points by outlier tests may not indicate that data are in any sense in error, but only that they do not follow a normal distribution (Fisher, 1922). For example, below are 25 observations from a lognormal distribution. When the t-prediction interval is applied with α=0.05, the largest observation is declared to be an outlier. Yet it is known to be from the same non-normal distribution as generated the remaining observations.

0.150	0.244	0.339	0.408	0.434
0.595	0.728	0.776	0.832	0.836
0.900	0.924	1.074	1.136	1.289
1.709	1.889	2.217	2.755	2.886
2.919	2.939	3.166	4.282	7.049

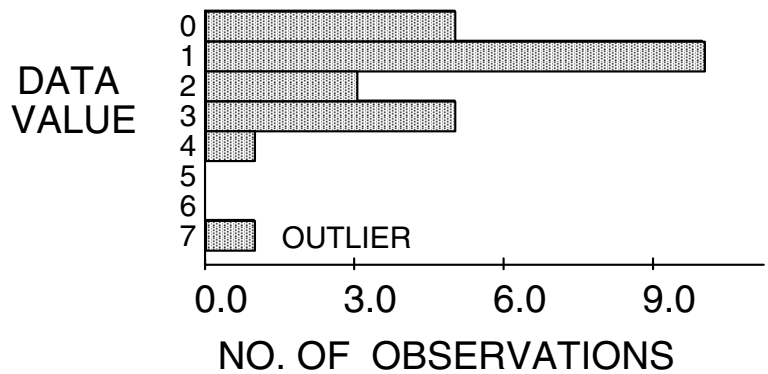


Table 3.3 Lognormal data set with "outlier" more than +2 sd above the mean.

Multiple outliers cause other problems for outlier tests that are based on normality (Beckman and Cook, 1983). They may so inflate the estimated standard deviation that no points are declared as outliers. When several points are spaced at increasingly larger distances from the mean, the first may be declared an outlier upon using the test once, but re-testing after deletion causes the second largest to be rejected, and so on. Replication of the test may eventually discard a substantial part of the data set. The choice of how many times to apply the test is entirely arbitrary.

3.8.2 Implications of Non-Normality for Quality Control

A visual presentation of confidence intervals used extensively in industrial processes is a **control chart** (Montgomery, 1991). A small number of products are sampled from the total possible at a given point in time, and their mean calculated. The sampling is repeated at regular or random intervals, depending on the design, resulting in a series of sample means. These are used to construct one type of control chart, the xbar chart. This chart visually detects when the mean of future samples become different from those used to construct the chart. The decision of

difference is based on exceeding the parametric confidence interval around the mean given in section 3.4.1.

Suppose a chemical laboratory measures the same standard solution at several times during a day to determine whether the equipment and operator are producing consistent results. For a series of n measurements at m time intervals, the total sample size N=n•m. The best estimate of the concentration for that standard is the overall mean

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

$\bar{X}$ is plotted as the center line of the chart. A confidence interval on that mean is described by equation 3.8, using the sample size n available for computing each individual mean value. Those interval boundaries are also plotted as parallel lines on the quality control chart. Mean values will on average plot outside of these boundaries only α•100% of the time if the means are normally distributed. Points falling outside the boundaries more frequently than this are taken to indicate that something in the process has changed.

If n is large (say 30 or more) the Central Limit Theorem states that the means will be normally distributed even though the underlying data may not be. However if n is much smaller, as is often the case, the means may not follow this pattern. In particular, for skewed data (data with outliers on only one side), the distribution around the mean may still be skewed. The result is a large value for the standard deviation, and wide confidence bands. Therefore the chart will have lower power to detect departures or drifts away from the expected mean value than if the data were not skewed.

Control charts are also produced to illustrate process variance. These either use the range (R chart) or standard deviation (S chart). Both charts are even more sensitive to departures from normality than is the $\bar{X}$ chart (Montgomery, 1991). Both will have a difficult time in detecting changes in variance when the underlying data are non-normal, and the sample size n for each mean is small.

In water quality studies the most frequent application of control charts is to laboratory chemical analyses. As chemical data tend to be positively skewed, control charts on the logs of the data are usually more applicable than those in the original units. Otherwise large numbers of samples must be used to determine mean values. Use of logarithms results in the center line estimating the median in original units, with multiplicative variation represented by the confidence bands of section 3.3.2.

Nonparametric control charts may be utilized if sample sizes are sufficiently large. These could use the confidence intervals for the median rather than the mean, as in section 3.3. Alternatively, limits could be set around the mean or median using the "F-psuedosigma" of Hoaglin (1983). This was done by Schroeder et al. (1987). The F-psuedosigma is the interquartile range divided by 1.349. It equals the standard deviation for a normal distribution, but is not as strongly affected by outliers. It is most useful for characterizing symmetric data containing outliers at both ends, providing a more resistant measure of spread than does the standard deviation.

3.8.3 Implications of Non-Normality for Sampling Design

The t-interval equations are also used to determine the number of samples necessary to estimate a mean with a specified level of precision. However, such equations require the data to approximately follow a normal distribution. They must consider power as well as the interval width. Finally, one must decide whether the mean is the most appropriate characteristic to measure for skewed data.

To estimate the sample size sufficient for determining an interval estimate of the mean with a specified width, equation 3.8 is solved for n to produce

$$n = \left(\frac{t_{\alpha/2,n-1}\, s}{\Delta} \right)^2 \qquad [3.20]$$

where s is the sample standard deviation and Δ is one-half the desired interval width. Sanders et al. (1983) and other authors have promoted this equation. As discussed above, for sample sizes less than 30 to 50 and even higher with strongly skewed data, this calculation may have large errors. Estimates of s will be inaccurate, and strongly inflated by any skewness and/or outliers. Resulting estimates of n will therefore be large. For example, Hakanson (1984) estimated the number of samples necessary to provide reasonable interval widths for mean river and lake sediment characteristics, including sediment chemistry. Based on the coefficients of variation reported in the article, the data for river sediments were quite skewed, as might be expected. Necessary sample sizes for rivers were calculated at 200 and higher.

Before using such simplistic equations, skewed data should be transformed to something closer to symmetry, if not normality. For example, logarithms will drastically lower estimated sample sizes for skewed data, equivalent to equation 3.13. Samples sizes would result which allow the median (geometric mean) to be estimated within a multiplicative tolerance factor equal to $\pm 2\Delta$ in log units.

A second problem with equations like 3.20 for estimating sample size, even when data follow a normal distribution, is pointed out by Kupper and Hafner (1989). They show that eq. 3.20 underestimates the true sample size needed for a given level of precision, even for estimates of $n \geq 40$. This is because eq. 3.20 does not recognize that the standard deviation s is only an

estimate of the true value σ. They suggest adding a **tolerance probability** to eq. 3.20, akin to a statement of power. Then the estimated interval width will be at least as small as the desired interval width for some stated percentage (say 90 or 95%) of the time. For example, when n would equal 40 based on equation 3.20, the resulting interval width will be less than the desired width 2Δ only about 42% of the time! The sample size should instead be 53 in order to insure the interval width is within tolerance range 90% of the time. They conclude that eq. 3.20 and similar equations which do not take power into consideration "behave so poorly in all instances that their future use should be strongly discouraged".

Sample sizes necessary for interval estimates of the median or to perform the nonparametric tests of later chapters may be derived without the assumption of normality required above for t-intervals. Noether (1987) describes these more robust sample size estimates, which do include power considerations and so are more valid than equation 3.20. However, neither the normal-theory or nonparametric estimates consider the important and frequently-observed effects of seasonality or trend, and so may never provide estimates sufficiently accurate to be anything more than a crude guide.

Exercises

3.1 Compute both nonparametric and parametric 95% interval estimates for the median of the granodiorite data of exercise 2.3. Which is more appropriate for these data? Why?

3.2 Compute the symmetric 95% interval estimate for the mean of the quartz monzonite data of exercise 2.3. Compute the sample mean, and the mean assuming the data are lognormal. Which point estimate is more appropriate for these data? Why?

3.3 A well yield of 0.85 gallons/min/foot was measured in a well in Virginia. Is this yield likely to belong to the same distribution as the data in exercise 1.1, or does it represent something larger? Answer by computing 95% parametric and nonparametric intervals. Which interval is more appropriate for these data?

3.4 Construct the most appropriate 95 percent interval estimates for the mean and median annual streamflows for the Conecuh River at Brantley, Alabama (data in Appendix C2).

3.5 Suppose a water intake is to be located on the Potomac River at Chain Bridge in such a way that the intake should not be above the water surface more than 10 percent of the time. Data for the design year (365 daily flows, ranked in order) are given in Appendix C3. Compute a 95% confidence interval for the daily flow guaranteed by this placement during the 90% of the time the intake is below water.

Chapter 4

Hypothesis Tests

Scientists collect data in order to learn about the processes and systems those data represent. Often they have prior ideas, called hypotheses, of how the systems behave. One of the primary purposes of collecting data is to test whether those hypotheses can be substantiated, with evidence provided by the data. Statistical tests are the most quantitative ways to determine whether hypotheses can be substantiated, or whether they must be modified or rejected outright.

One important use of hypothesis tests is to evaluate and compare groups of data. Water resources scientists have made such comparisons for years, sometimes without formal test procedures. For example, water quality has been compared between two or more aquifers, and some statements made as to which are different. Historic frequencies of exceeding some critical surface-water discharge have been compared with those observed over the most recent 10 years. Rather than using hypothesis tests, the results are sometimes expressed as the author's educated opinions -- "it is clear that development has increased well yield." Hypothesis tests have at least two advantages over educated opinion:

1) they insure that every analyst of a data set using the same methods will arrive at the same result. Computations can be checked on and agreed to by others.
2) they present a measure of the strength of the evidence (the p-value). The decision to reject an hypothesis is augmented by the risk of that decision being incorrect.

In this chapter hypothesis tests are classified based on when each is appropriate for use. The basic structure of hypothesis testing is introduced. The rank-sum test is used to illustrate this structure, as well as to illustrate the origin of tables of test statistic quantiles found in most statistics textbooks. Finally, tests for normality are discussed. Concepts and terminology found here will be used throughout the rest of the book.

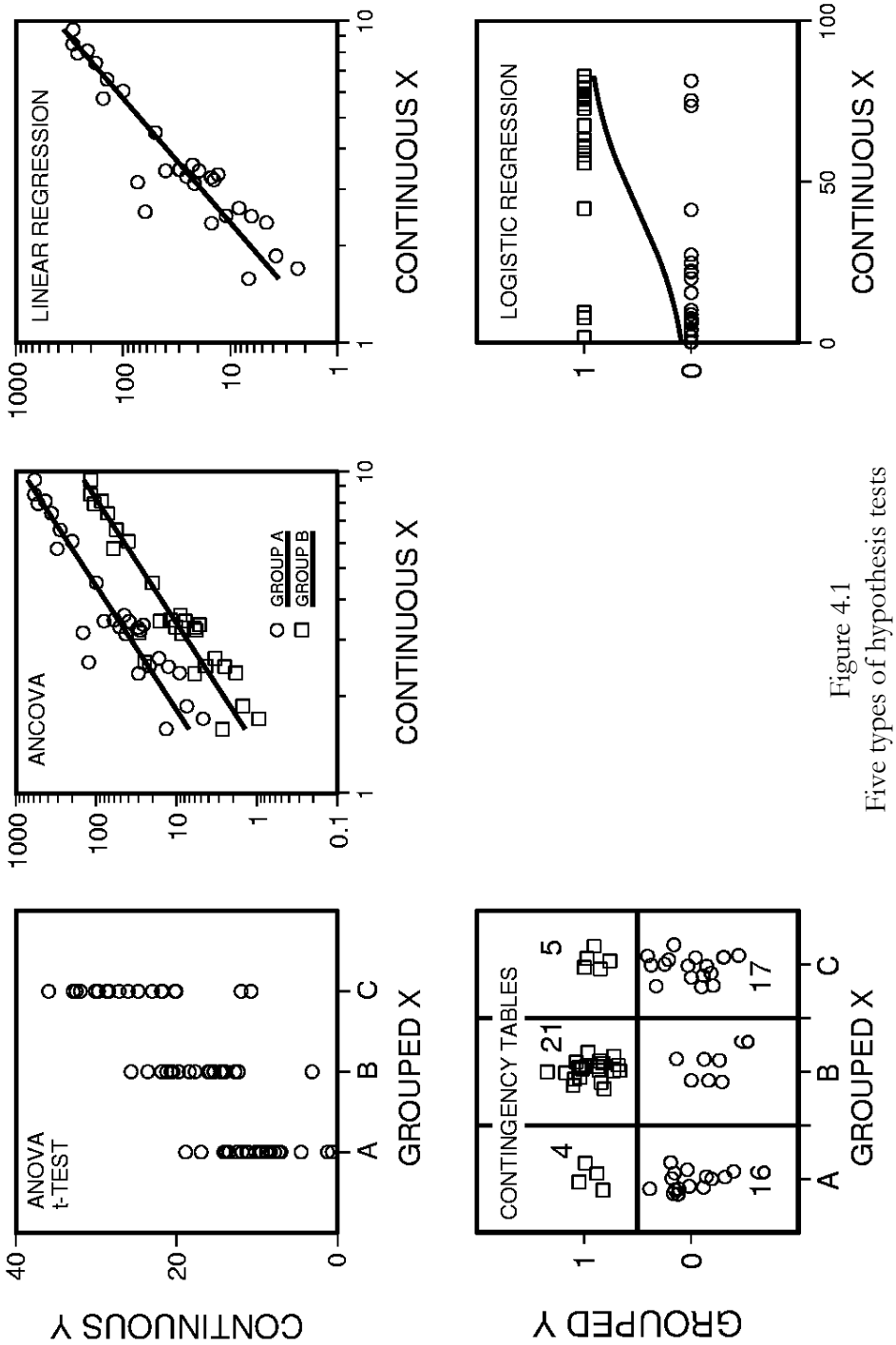


Figure 4.1
Five types of hypothesis tests

4.1 Classification of Hypothesis Tests

The numerous varieties of hypothesis tests often cause unnecessary confusion to scientists. Tests can be classified into the five major types shown in figure 4.1, based on the measurement scales of the data being tested. Within these types, the distributional shape of the data determine which of two major divisions of hypothesis tests, parametric or nonparametric, are appropriate for use. Thus the data, along with the objectives of the study, determine which test procedure should be employed.

The terms **response variable** and **explanatory variable** are used in the following discussion. A response variable is one whose variation is being studied. In the case of regression, for example, the response variable is sometimes called the "dependent variable" or "y variable". An explanatory variable is one used to explain why and how the magnitude of the response variable changes. With a t-test, for example, the explanatory variable consists of the two categories of data being tested.

4.1.1 Classification Based on Measurement Scales

In figure 4.1, five groupings of test procedures are represented by the five boxes. Each differs only in the measurement scales of the response and explanatory variables under study. The scales of measurement may be either continuous or categorical. Both parametric and nonparametric tests may be found within a given box.

Tests represented by the three boxes in the top row of figure 4.1 are all similar in that the response variable is measured on a continuous scale. Examples of variables having a continuous scale are concentration, streamflow, porosity, and many of the other items measured by water resources scientists. Tests represented by the two boxes along the bottom of figure 4.1, in contrast, have response variables measured only on a categorical or grouped measurement scale. These variables can only take on a finite, usually small, number of values. They are often designated as letters or integer values. Categorical variables used primarily as explanatory variables include aquifer type, month, land use group, and station number. Categorical variables used as response variables include above/below a reporting limit (perhaps recorded as 0 or 1), presence or absence of a particular species, and low/medium/high risk of contamination.

The top left box represents the two- and multi-sample hypothesis tests such as the rank-sum and t-tests. The subject of Chapters 5 through 7, these tests determine whether a continuous response variable (such as concentration) differs in its central value among two or more grouped explanatory variables (such as aquifer unit).

The top right box represents two often-used methods -- linear regression and correlation. Both relate a continuous response variable (the dependent or y variable) to a continuous explanatory variable (the independent or x variable). Examples include regression of the 100-year flood

magnitude versus basin characteristics, and correlations between concentrations of two chemical constituents. Analysis of trends over time is a special case of this class of methods, where the explanatory variable of primary interest is time.

The top center box is a blend of these two approaches, called analysis of covariance. A continuous response variable is related to several explanatory variables, some of which are continuous and some categorical. This is discussed in Chapter 11.

The bottom left box represents a situation similar to that for use of t-tests or analysis of variance, except that the response variable is categorical. Examples include determining whether the probability of finding a volatile organic above the reporting limit varies by land-use grouping. Contingency tables appropriately measure the association between two such categorical variables. Further information is found in Chapter 14.

The bottom right box shows that a regression-type relationship can be developed for the case of a categorical response variable. Perhaps the proportion of pesticide or other data below the reporting limit exceeds fifty percent, and it makes little sense to try to model mean or median concentrations. Instead, the probability of finding a detectable concentration can be related to continuous variables such as population density, percent of impervious surface, irrigation intensities, etc. This is done through the use of logistic regression, one subject of Chapter 15. Logistic regression can also incorporate categorical explanatory variables in a multiple regression context, making it the equivalent of analysis of covariance for categorical response variables.

4.1.2 Classification Based on the Data Distribution

Hypothesis tests which assume that the data have a particular distribution (usually a normal distribution, as in Fig. 1.2) are called **parametric tests**. This is because the information contained in the data is summarized by parameters, usually the mean and standard deviation, and the test statistic is computed using these parameters. This is an efficient process if the data truly follow the assumed distribution. When they do not, however, the parameters may only poorly represent what is actually occurring in the data. The resulting test can then reach an incorrect conclusion, usually because it lacks sensitivity (power) to detect real effects.

Hypothesis tests not requiring the assumption that data follow a particular distribution are called distribution-free or **nonparametric tests**. Information is extracted from the data by comparing each value with all others (ranking the data) rather than by computing parameters. A common misconception is that nonparametric tests "lose information" in comparison to parametric tests because nonparametric tests "discard" the data values. Bradley (1968, p.13) responded to this misconception: "Actually, the utilization of the additional sample information [in the parameters] is made possible by the additional population 'information' embodied in the parametric test's assumptions. Therefore, the distribution-free test is discarding information only if the parametric test's assumptions are known to be true." Rather than discarding

information, nonparametric tests efficiently extract information on the relative magnitudes (ranks) of data without collapsing the information into only a few simple statistics. Both parametric and nonparametric tests will be presented in the upcoming chapters for each category of hypothesis tests.

4.2 Structure of Hypothesis Tests

Hypothesis tests are performed by following the structure discussed in the next six sections:

STRUCTURE OF HYPOTHESIS TESTS

1) Choose the appropriate test.
2) Establish the null and alternate hypotheses.
3) Decide on an acceptable error rate α.
4) Compute the test statistic from the data.
5) Compute the p-value.
6) Reject the null hypothesis if $p \leq \alpha$.

4.2.1 Choose the Appropriate Test

Test procedures are selected based on the data characteristics and study objectives. Figure 4.1 presented the first selection criteria -- the measurement scales of the data. The second criteria is the objective of the test. Hypothesis tests are available to detect differences between central values of two groups, three or more groups, between spreads of data groups, and for covariance between two or more variables, among others. For example, to compare central values of two independent groups of data, either the t-test or rank-sum test might be selected (see figure 4.2). Subsequent chapters are organized by test objectives, with several alternate tests discussed in each.

The third selection criteria is the choice between parametric or nonparametric tests. This should be based on the expected distribution of the data involved. If similar data in the past were normally distributed, a parametric procedure would usually be selected. If data were expected to be non-normal, or not enough is known to assume any specific distribution, nonparametric tests would be preferred. The power of parametric tests to reject H_0 when H_0 is false can be quite low when applied to non-normal data, and type II errors commonly result (Bradley, 1968). This loss of power is the primary concern when using parametric tests.

Sometimes the choice of test is based on a prior test of normality for that particular data set. If normality is rejected a nonparametric test is chosen. Otherwise, a parametric test is used. This can lead to two problems. First, with small data sets it is difficult to reject the null hypothesis of normality because there is so little evidence on which to base a decision. Tests based on little

data have little power. Thus a parametric test might easily be used when the underlying data are actually non-normal. Nonparametric tests are particularly appropriate for small data sets unless experience supports the assumption of normality. Second, small departures from normality not large enough to detect with a test may be sufficiently large to weaken the power of parametric tests. An example is given in Chapter 10. For nearly-normal data, such as produced by power transformations to near-symmetry, the two classes of methods will often give the same result.

Test procedures should be selected that have greater power for the types of data expected to be encountered. Comparisons of the power of two test procedures, one parametric and one nonparametric, can be based on the tests' asymptotic relative efficiencies (ARE), a property of their behavior with large sample sizes (Bradley, 1968, p.58). A test with larger ARE will have generally greater power. For non-normal data the ARE of nonparametric tests can be many times those of parametric tests(Hollander and Wolfe, 1973). Thus their power to reject H_0 when it is truly false is generally much higher in this case. When data are produced by a normal distribution, nonparametric tests have generally lower (5-15%) ARE than parametric tests (Hollander and Wolfe, 1973). Thus nonparametric tests are, in general, never much worse than their parametric counterparts in their ability to detect departures from the null hypothesis, and may be far, far better. As an example, the rank-sum test has a larger ARE (more power) than the t-test for distributions containing outliers (Conover, 1980, p.225). Kendall and Stuart (1979, p.540) show that for the gamma distribution (a skewed distribution commonly used in water resources) a moderate skew of 1.15 produces an ARE of greater than 1.25 for the rank-sum versus the t test. As skewness increases, so does the ARE. Therefore in the presence of skewness and outliers, precisely the characteristics commonly shown by water resources data, nonparametric tests exhibit greater power than do parametric tests.

One question which always arises is how non-normal must a distribution be in order for nonparametric tests to be preferred? Blair and Higgins (1980) gave insight into this question. They mixed data from two normal distributions, 95 percent from one normal distribution and 5 percent from a second normal distribution with quite different mean and standard deviation. Such a situation could easily be envisioned when data result from low to moderate discharges with occasional storm events, or from a series of wells where 5 percent are affected by a contaminant plume, etc. A difference of 5 percent from truly normal may not be detectable by a graph or test for normality. Yet when comparing two groups of this type, they found that the rank-sum test exhibited large advantages in power over the t-test. As a result, data groups correctly discerned as different by the rank-sum test could be found "not significantly different" by the t-test. Their paper is recommended for further detail and study.

The greatest strengths of parametric procedures are in modeling and estimation, such as performed with regression. Relationships among multiple variables can be described and tested which are difficult, if not nearly impossible, with nonparametric methods. Statistical practice has

historically been dominated by parametric procedures, due largely to their computational elegance.

Transformations are sometimes used to make data more normally distributed, prior to performing a parametric test. There is no guarantee that a given transformation, such as taking logarithms, will produce data sufficiently close to a normal distribution. Often several attempts to find a suitable transformation are required before the data appear approximately normal. The primary pitfall in using transformations is that when two or more groups are to be compared, no single transformation may provide nearly-normal data simultaneously for all groups. Groups whose right-skewness was solved by transformation may be offset by relatively symmetric groups which are now left-skewed. When several tests are performed, such as trend tests at numerous locations, parametric tests might be appropriate in some cases but not in others. Comparisons of results across sites are more difficult when test procedures and/or transformations vary for each case. Nonparametric tests allow the freedom to use the identical test procedure in all cases, without the requirement that the many individual data sets follow the same distribution. Finally, transformations may produce nearly-symmetric data, but cannot compensate for a heavy-tailed distribution -- the presence of more data near the extremes than found in a normal distribution.

It should be noted that there are actually three versions of most nonparametric tests:

1. **Exact test**. Exact versions of nonparametric tests provide results (in the form of p-values, defined soon) which are exactly correct . They are computed by comparing the test statistic to a table of quantiles that is specific for the sample sizes present. Therefore an extensive set of tables is required, one for every possible combination of sample sizes. When sample sizes are small, only the exact version will provide accurate results.
2. **Large sample approximation**. To avoid the necessity for large books filled with tables of test statistic quantiles, approximate p-values are obtained by assuming that the distribution of the test statistic can be approximated by some common distribution, such as the normal. This does not mean the data themselves follow that distribution, but only that the test statistic does. For large sample sizes (30 or more observations per group, but sometimes less) this approximation is very accurate. The test statistic is modified if necessary (often standardized by subtracting its mean, and dividing by its standard deviation), and then compared to a table of the common distribution to determine the p-value.
   <u>WARNING</u>: Computer software predominantly uses large sample approximations when reporting p-values, whether or not the sample sizes are sufficient to warrant using them. For small sample sizes, p-values should be taken from exact tables rather than from the computer printout.
3. **Rank transformation test**. In this approximation, parametric procedures are computed not on the data themselves, but on the ranks of the data (smallest observation has rank=1, largest has rank=N). Conover and Iman (1981) have shown this to adequately approximate many exact nonparametric tests for large samples sizes. The rank-sum test would be

approximated in this fashion by computing a t-test on joint ranks of the data. In fact, Iman and Conover (1983) use the name "rank-sum test" for just this procedure. We would call this version a
"t-test on ranks", reserving the traditional name for the first or second versions of the test and more accurately describing what was done. Rank approximations are most useful when performing nonparametric tests using statistics packages which contain only parametric procedures. They are also very useful for situations where there is no equivalent nonparametric analog, such as for multiple-factor analysis of variance.

In figure 4.2, exact and rank transform tests are aligned with their parametric counterparts, as a guide to the use of hypothesis tests.

4.2.2 Establish the Null and Alternate Hypotheses

The null and alternate hypotheses should be established prior to collecting data. These hypotheses are a concise summary of the study objectives, and will keep those objectives in focus during data collection.

The **null hypothesis** (H_0) **is what is assumed to be true about the system under study prior to data collection, until indicated otherwise**. It usually states the "null" situation -- no difference between groups, no relation between variables. One may "suspect", "hope", or "root for" either the null or alternate hypothesis, depending on one's vantage point. But the null hypothesis is what is assumed true until the data indicate that it is likely to be false. For example, an engineer may test the hypothesis that wells upgradient and downgradient of a hazardous waste site have the same concentrations of some contaminant. They may "hope" that downgradient concentrations are higher (the company gets a new remediation project), or that they are the same (the company did the original site design!). In either case, the null hypothesis assumed to be true is the same: concentrations are similar in both groups of wells.

The **alternate hypothesis** (H_1) **is the situation anticipated to be true if the evidence (the data) show that the null hypothesis is unlikely**. It is in some cases just the negation of H_0, such as "the 100-year flood is not equal to the design value." H_1 may also be more specific than just the negation of H_0 -- "the 100-year flood is greater than the design value." Alternate hypotheses come in two general types: one-sided, and two-sided. Their associated hypothesis tests are called one-sided and two-sided tests. These are often confused and misapplied.

Two-sided tests occur when evidence in either direction from the null hypothesis (larger or smaller, positive or negative) **would cause the null hypothesis to be rejected** in favor of the alternate hypothesis. For example, if evidence that "the 100-year flood is smaller than the design value" <u>or</u> "the 100-year flood is greater than the design value" would both cause doubt about the null hypothesis, the test is two-sided. Most tests in water resources are of this kind.

PARAMETRIC	NONPARAMETRIC [exact]	RANK TRANSFORM [approximation]
Two Independent Data Groups (Chapter 5)		
two-sample t-test	rank sum test *or* Mann-Whitney *or* Wilcoxon-Mann-Whitney	t-test on ranks
Matched Pairs of Data (Chapter 6)		
paired t-test	(Wilcoxon) signed-rank test	t-test on signed ranks
More than Two Independent Data Groups (Chapter 7)		
1-way Analysis Of Variance (ANOVA)	Kruskal-Wallis test	1-way ANOVA on ranks
More than Two Dependent Data Groups (Chapter 7)		
Analysis Of Variance without replication	Friedman test	2-way ANOVA on ranks
Correlation between Two Continuous Variables (Chapter 8)		
Pearson's r *or* linear correlation	Kendall 's tau	Spearman's rho (Pearson's r on ranks)
Relation between Two Continuous Variables (Chapters 9, 10)		
Linear Regression test for slope = 0	Mann-Kendall test for slope = 0	regression on ranks: test for monotonic change

Figure 4.2 Guide to the classification of some hypothesis tests

One-sided tests occur when departures in only one direction from the null hypothesis **would cause the null hypothesis to be rejected** in favor of the alternate hypothesis. With

one-sided tests, it is considered supporting evidence for H_0 should the data indicate differences opposite in direction to the alternate hypothesis. For example, suppose only evidence that the 100-year flood is greater than the previous design value is of interest, as only then must the culvert be replaced. The null hypothesis would be stated as "the 100-year flood is less-than or equal to the design flood", while the alternate hypothesis is that "the 100-year flood exceeds the design value." Any evidence that the 100-year flood is smaller than the design value is considered evidence for H_0.

If it cannot be stated prior to looking at any data that departures from H_0 in only one direction are of interest, a two-sided test should be performed. If one simply wants to look for differences between two streams or two aquifers or two time periods, then a two-sided test is appropriate. It is not appropriate to look at the data, find that group A is considerably larger in value than group B, and perform a one-sided test that group A is larger. This would be ignoring the real possibility that had group B been larger there would have been interest in that situation as well. Examples in water resources where one-sided tests would be appropriate are:

1. testing for decreased annual floods or downstream sediment loads after completion of a flood-control dam,
2. testing for decreased nutrient loads or concentrations due to a new sewage treatment plant or best management practice,
3. testing for an increase in concentration when comparing a suspected contaminated site to an upstream or upgradient control site.

4.2.3 Decide on an Acceptable Error Rate α

The α-value, or significance level, is the probability of incorrectly rejecting the null hypothesis (rejecting H_0 when it is in fact true, called a "Type I error"). Figure 4.3 shows that this is one of four possible outcomes of an hypothesis test. The significance level is the risk of a Type I error deemed acceptable by the decision maker. It is a "management tool" dependent not on the data, but on the objectives of the study. Statistical tradition uses a default of 5% (0.05) for α, but there is no reason why other values should not be used. Suppose that an expensive cleanup process will be mandated if the null hypothesis of "no contamination" is rejected, for example. The α-level for this test might be set very small (such as 1%) in order to minimize the chance of needless cleanup costs. On the other hand, suppose the test was simply a first cut at classifying sites into "high" and "low" values prior to further analysis of the "high" sites. In this case the α-level might be set to 0.10 or 0.20, so that all sites with high values would likely be retained for further study.

		Unknown True Situation	
		H_0 is true	H_0 is false
Decision	Fail to Reject H_0	Correct decision Prob(correct decision) = $1-\alpha$	Type II error Prob(Type II error) = β
	Reject H_0	Type I error Prob (Type I error) = α **Significance level**	Correct decision Prob (correct decision) = $1-\beta$ **Power**

Figure 4.3 Four possible results of hypothesis testing.

Since α represents one type of error, why not keep it as small as possible? One way to do this would be to never reject H_0 -- α would then equal zero. Unfortunately this would lead to large errors of a second type -- failing to reject H_0 when it was in fact false. This second type of error is called a Type II error, or lack of power (Fig. 4.3). Both errors are of concern to practitioners, and both will have some finite probability of occurrence unless decisions to "always reject" or "never reject" are made. Once a decision is made as to an acceptable Type I risk α, two steps can be taken to concurrently reduce the risk of Type II error β:

1. Increase the sample size n.
2. Use the test procedure with the greatest power for the type of data being analyzed.

For water quality applications, null hypotheses are usually of "no contamination". Situations with low power mean that actual contamination may not be detected. This happens with simplistic formulas for determining sample sizes (Kupper and Hafner, 1989). Instead, probabilities of Type II errors should be considered when setting sample size. Power is also sacrificed when data having the characteristics outlined in Chapter 1 are analyzed with tests requiring a normal distribution. Power loss increases as skewness and the number of outliers increase.

4.2.4 Compute the Test Statistic from the Data

Test statistics summarize the information contained in the data. If the test statistic is not unusually different from what is expected to occur if the null hypothesis is true, the null hypothesis is not rejected. However, if the test statistic is a value unlikely to occur when H_0 is true, the null hypothesis is rejected. The p-value measures how unlikely the test statistic is when H_0 is true.

4.2.5 Compute the p-Value

The p-value is the probability of obtaining the computed test statistic, or one even less likely, when the null hypothesis is true. It is derived from the data, concisely expressing the evidence against the null hypothesis contained in the data. It measures the "believability" of the null hypothesis. The smaller the p-value, the less likely is the observed test statistic when H_0 is true, and the stronger the evidence for rejection of the null hypothesis. The p-value is also called the "attained significance level", the significance level attained by the data.

How do p-values differ from α levels? The α-level does not depend on the data, but states the risk of making a Type I error that is acceptable *a priori* to the scientist or manager. The α-level is the critical value which allows a "yes/no" decision to be made -- the treatment plant has improved water quality, nitrate concentrations in the well exceed standards, etc.. The p-value provides more information -- the strength of the scientific evidence. Reporting the p-value allows someone with a different risk tolerance (different α) to make their own yes/no decision.

For example, consider a test of whether upgradient and downgradient wells have the same expected contaminant concentrations. If downgradient wells show evidence of higher concentrations, some form of remediation will be required. Data are collected, and a test statistic calculated. A decision to reject at α=0.01 is a statement that "remediation is warranted as long as there is less than a 1 percent chance that the observed data would occur when upgradient and downgradient wells actually had the same concentration." This level of risk was settled on as acceptable, so that 1 percent of the time remediation would be performed when in fact it is not required. Reporting only "reject" or "not reject" would prevent the audience from distinguishing a case that is barely able to reject (p=0.009) from one in which H_0 is virtually certain to be untrue (p=0.0001). Reporting a p-value of 0.02, for example, would allow a later decision by someone with a greater tolerance of unnecessary cleanup (α = 5 percent, perhaps) to decide for remediation.

4.2.6 Make the Decision to Reject H_0 or Not

Reject H_0 when: p-value < α-level.

When the p-value is less than the decision criteria (the α-level), H_0 is **rejected**. When the p-value is greater than α, H_0 is **not rejected**. The null hypothesis is never "accepted", or proven to be true. It is <u>assumed</u> to be true until proven otherwise, and is "not rejected" when there is insufficient evidence to do so.

4.3 The Rank-Sum Test as an Example of Hypothesis Testing

Suppose that aquifers X and Y are sampled to determine whether the concentrations of a contaminant in the aquifers are similar or different. This is a test for differences in location or central value, and will be covered in detail in Chapter 5. Two samples x_i are taken from aquifer X (n=2), and 5 samples y_i from aquifer Y (m=5) for a total of 7 samples (N = n+m = 7). Also suppose that there is a prior reason to believe that X values tend to be lower than Y values: aquifer X is deeper, and is likely to be uncontaminated. The null hypothesis (H_0) and alternative hypothesis (H_1) of this one-sided test are as follows:

H_0: x_i and y_i are samples from the same distribution, or
H_0: Prob $(x_i \geq y_i) = 0.5$.

H_1: x_i is from a distribution which is generally lower that of y_i, or
H_1: Prob $(x_i \geq y_i) < 0.5$.

Remember that with one-sided tests such as this one, data indicating differences opposite in direction to H_1 (x_i frequently larger than y_i) are considered supporting evidence for H_0. With one-sided tests we can only be interested in departures from H_0 in one direction.

Having established the null and alternate hypotheses, an acceptable error rate α must be set. As in a court of law, innocence is assumed (i.e. concentrations are identical) unless evidence is collected to show "beyond a reasonable doubt" that aquifer Y has higher concentrations (i.e. that differences observed are not likely to have occurred by chance alone). The "reasonable doubt" is set by α, the significance level.

If the t-test were to be considered as the test procedure, each data group should be tested for normality. However, sample sizes of 2 and 5 are too small for a reliable test of normality. Thus the nonparametric rank-sum test is appropriate. This test procedure entails ranking all 7 values (lowest concentration has rank=1, highest has rank=7) and summing the ranks of the 2 values from the population with the smaller sample size (X). This rank-sum is the statistic W used in the exact test.

Next, W would be computed and compared to a table of test statistic quantiles to determine the p-value. Where do these tables come from? We will derive the table for sample sizes 2 and 5 as an example.

What are the possible values W may take, given that the null hypothesis is true? The collection of all of the possible outcomes of W defines its distribution, and therefore composes the table of rank-sum test statistic quantiles. Shown below are all the possible combinations of ranks of the two x values.

1,2	1,3	1,4	1,5	1,6	1,7
	2,3	2,4	2,5	2,6	2,7
		3,4	3,5	3,6	3,7
			4,5	4,6	4,7
				5,6	5,7
					6,7

If H_0 is true, each of the 21 possible outcomes must be equally likely. That is, it is just as likely for the two x's to be ranks 1 and 2, or 3 and 5, or 1 and 7, etc. Each one of the outcomes results in a value of W, the sum of the two ranks. The 21 W values corresponding to the above outcomes are

3	4	5	6	7	8
	5	6	7	8	9
		7	8	9	10
			9	10	11
				11	12
					13

The expected value of W is the mean (and median) of the above values, or 8. Given that each outcome is equally likely when H_0 is true, the probability of each possible W value is:

W	3	4	5	6	7	8	9	10	11	12	13
Prob(W)	1/21	1/21	2/21	2/21	3/21	3/21	3/21	2/21	2/21	1/21	1/21

What if the data collected produced 2 x values having ranks 1 and 4? Then W would be 5, lower than the expected value E [W] = 8. If H_1 were true rather than H_0, W would tend toward low values. What is the probability that W would be as low as 5 or lower if H_0 were true? It is the sum of the probabilities for W = 3, 4, and 5, or 4/21 = 0.190 (see figure 4.4). **This number is the p-value for the test statistic of 5**. It says that the chance of a departure from E [W] of at least this magnitude occurring when H_0 is true is 0.190, which is not very uncommon (about 1 chance in 5). Thus the evidence against H_0 is not too convincing. If the ranks of the 2 x values had been 1 and 2, then W = 3 and the p-value would be 1/21 = 0.048. This result is much less likely than the previous case but is still not extremely rare. In fact, due to such a small sample size the test can never result in a highly compelling case for rejecting H_0. Adding more data would make it possible to attain lower p-values, providing a stronger case against H_0.

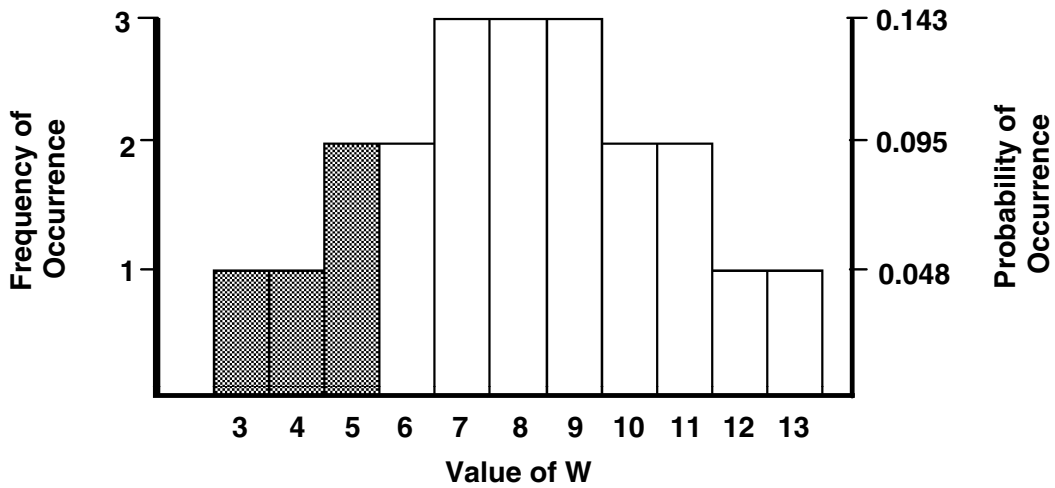


Figure 4.4 Probabilities of occurrence for a rank-sum test with sample sizes of 2 and 5. The p-value for a one-sided test equals the area shaded.

This example has considered only the one-sided p-value, which is appropriate when there is some prior notion that x should be smaller than y (or the reverse). Quite often the situation is that there is no prior notion of which should be lower. In this case a two-sided test must be done. The two-sided test has the same null hypothesis as was stated above, but H_1 is now that x_i and y_i are from different distributions, or

H_1: Prob $(x_i \geq y_i) \neq 0.5$.

Suppose that W for the two-sided test were found to be 5. The p-value equals the probability that W will differ from E [W] by this much or more, in either direction. It is

Prob $(W \leq 5)$ + Prob $(W \geq 11)$. (see figure 4.5)

Where did the 11 come from? It is just as far from E [W] = 8 as is 5. The two-sided p-value therefore equals 8/21 = 0.381, twice the one-sided p-value. Symbolically we could state:

Prob $(|W - E[W]| \geq 3) = 8/21$.

To summarize the subject of p-values: they describe how "far" the observed test statistic is from that expected to occur if the null hypothesis were true. They are the probability of being that far or farther given that the null hypothesis is true. **The lower the p-value the stronger is the case against the null hypothesis.**

Now, lets look at an α-level approach. Return to the original problem, the case of a one-sided test. Assume α is set equal to 0.1. This corresponds to a critical value for W, call it W*, such that Prob $(W \leq W^*) = \alpha$. Whenever $W \leq W^*$, H_0 is rejected with no more than a 0.1 frequency of error if H_0 were always true. However, because W can only take on discrete, in fact integer, values as seen above, a W* which exactly satisfies the equation is not usually available. Instead the largest possible W* such that Prob $(W \leq W^*) \leq \alpha$ is used.

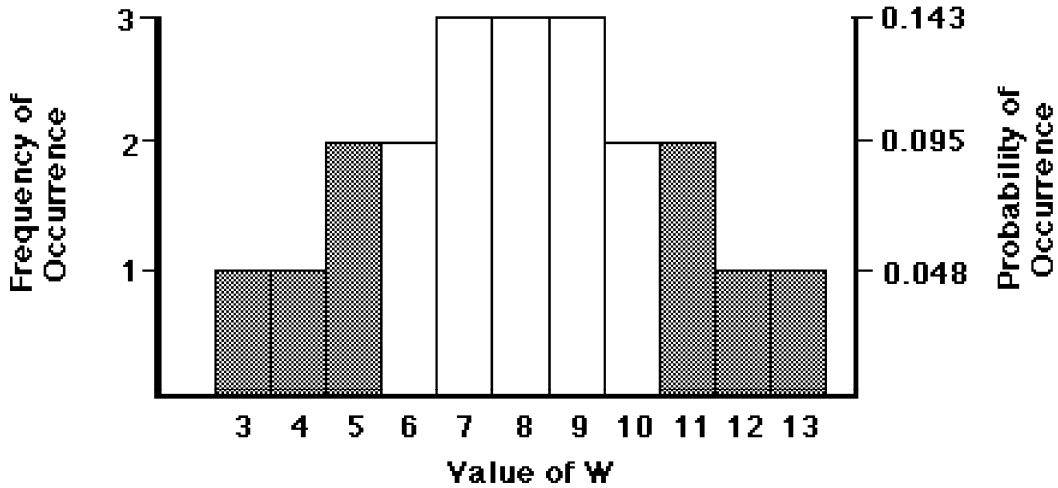


Figure 4.5 Probabilities of occurrence for a rank-sum test with sample sizes of 2 and 5. The p-value for a two sided-test equals the area shaded.

Searching the above table of possible W values and their probabilities, W* = 4 because Prob (W ≤ 4) = 0.095 ≤ 0.1. Note the "lumpiness" of the relationship between α and W*. If α =0.09 had been selected then W* would be 3. This lumpiness can be avoided by reporting p-values rather than only "reject" or "not reject".

For a two-sided test a pair of critical values W_U^* and W_L^* is needed, where

$$\text{Prob}\ (W \le W_L^*) + \text{Prob}\ (W \ge W_U^*) \le \alpha \quad \text{and} \quad W_U^* - E\,[W] = E\,[W] - W_L^*.$$

These upper and lower critical values of W are symmetrical around E [W] such that the probability of W falling on or outside of these critical levels is as close as possible to α, without exceeding it, under the assumption that H_0 is true. In the case at hand, if α = 0.1, then W_L^*= 3 and W_U^*= 13 because

$$\text{Prob}\ (W \le 3) + \text{Prob}\ (W \ge 13) = 0.048 + 0.048 = 0.095 \le 0.1.$$

Note that for a two-sided test, the critical values are farther from the expected value than in a one-sided test at the same α level.

It should be recognized that p-values are also influenced by sample size. For a given magnitude of difference between the x and y data, and a given amount of variability in the data, p values will tend to be smaller when the sample size is large. In the extreme case where vast amounts of data are available, it is a virtual certainty that p values will be small even if the differences between x and y are what might be called "of no practical significance."

Most statistical tables are set up for one-sided tests. That is, the rejection region α or the p-value is given in only one direction. **When a two-sided test at significance level α is performed, the tables must be entered using $\alpha/2$.** In this way rejection can occur with a

probability of $\alpha/2$ on either side, and an overall probability of α. **Similarly, tabled p-values must be doubled to get p-values for a two-sided test.** Modern statistical software often reports p-values with its output, eliminating the need for tables. Be sure to know whether it is one-sided or two-sided p-values being reported.

4.4 Tests for Normality

The primary reason to test whether data follow a normal distribution is to determine if parametric test procedures may be employed. The null hypothesis for all tests of normality is that the data are normally distributed. Rejection of H_0 says that this is doubtful. Failure to reject H_0, however, does not prove that the data do follow a normal distribution, especially for small sample sizes. It simply says normality cannot be rejected with the evidence at hand. Use of a larger α-level (say 0.1) will increase the power to detect non-normality, especially for small sample sizes, and is recommended when testing for normality.

The test for normality used in this book is the probability plot correlation coefficient (PPCC) test discussed by Looney and Gulledge (1985a). Remember from Chapter 2 that the more normal a data set is, the closer it plots to a straight line on a normal probability plot. To test for normality, this linearity is tested by computing the linear correlation coefficient between data and their normal quantiles (or "normal scores", the linear scale on a probability plot). Samples from a normal distribution will have a correlation coefficient very close to 1.0. As data depart from normality, their correlation coefficient will decrease below 1. To perform a test of H_0: the data are normal versus H_1: they are not, the correlation coefficient (r) between the data and their normal quantiles is tested to see if it is significantly less than 1. For a sample size of n, if r is smaller than the critical value r* of table B3 for the desired α-level, reject H_0. Looney and Gulledge (1985b) have shown this table, developed using the Blom plotting position, is also valid for other plotting positions except the Weibull position i/(n+1). In order to use one plotting position for all functions in this book, the Cunnane plotting position was adopted as explained in Chapter 2.

To illustrate this test, probability plots of the unit well yield data from Chapter 2 are shown in figures 4.6 and 4.7. For the valleys without fracturing, r = 0.805, the correlation coefficient between y_i and Zp in the left-hand side of Table 4.1.

From table B3 with n=12, if r is below the α = 0.05 critical value of r* = .928, normality is rejected. Therefore normality is rejected for the yields without fractures at α = 0.05. A p-value for this test would be <0.005, as r=0.805 is less than the tabled r* of 0.876 for α=0.005. Note the nonlinearity of the data on the probability plot (figure 4.6). For the yields with fracturing, n=13, r* is 0.932 at α = 0.05, and the PPCC r = 0.943; therefore fail to reject normality at α=0.05. The p-value for the yields with fracturing is just under 0.10 (normality would barely be

rejected at α=0.10). The probability plot, figure 4.7, shows a closer adherence to a straight line than for figure 4.6.

Table 4.1. Unit well yields (in gal/min/ft) in Virginia (Wright, 1985)
y_i = yield Zp = normal quantile

valleys without fracturing						valleys with fracturing					
y_i	Zp	y_i	Zp	y_i	Zp	y_i	Zp	y_i	Zp	y_i	Zp
0.001	−1.65	0.030	−.31	0.10	.52	0.020	−1.69	0.16	−.39	0.40	.39
0.003	−1.13	0.040	−.10	0.454	.80	0.031	−1.17	0.16	−.19	0.44	.60
0.007	−0.80	0.041	.10	0.49	1.13	0.086	−0.85	0.18	.00	0.51	.85
0.020	−0.52	0.077	.31	1.02	1.65	0.13	−0.60	0.30	.19	0.72	1.17
										0.95	1.69

Computer packages use several methods for testing normality. Several are based on probability plots. The most common is perhaps the Shapiro-Wilk test, as its power to detect non-normality is as good or better than other tests (Shapiro et al., 1968). A table of quantiles for this test statistic is available for n < 50 (Conover, 1980). Shapiro and Francia (1972) developed a modification of the Shapiro-Wilk test useful for all sample sizes. It is essentially identical to the PPCC test, as it is the r^2 for a regression between the data and their normal scores. Therefore p-values and power characteristics for the two tests should be essentially the same.

Tests for normality not related to probability plots include the Kolmogorov and chi-square tests, described in more detail by Conover (1980). Both are general tests that may be used for data which are ordinal (data recorded only as low/medium/high, etc) but do not possess a continuous scale. This makes them less powerful than the probability plot tests, however, for the specific purpose of testing continuous data for normality (Shapiro et al., 1968).

The important advantage of the PPCC test is its graphical analog, the probability plot, which visually illustrates its results. The probability plot itself provides information on how the data depart from normality, something not provided by any test statistic.

To make the PPCC test easy to perform by hand, normal quantiles for the Cunnane plotting positions of table B1 are listed in table B2 of the Appendix. For the n=12 yields without fracturing, for example, the upper six quantiles are easily found in the table. Lower quantiles are mirror images around zero of the upper quantiles, and so equal the upper values multiplied by −1. Table B2 quantiles were computed by first calculating the Cunnane plotting position to more significant digits than found in table B1, and then looking up the corresponding normal quantiles in a table of the normal distribution.

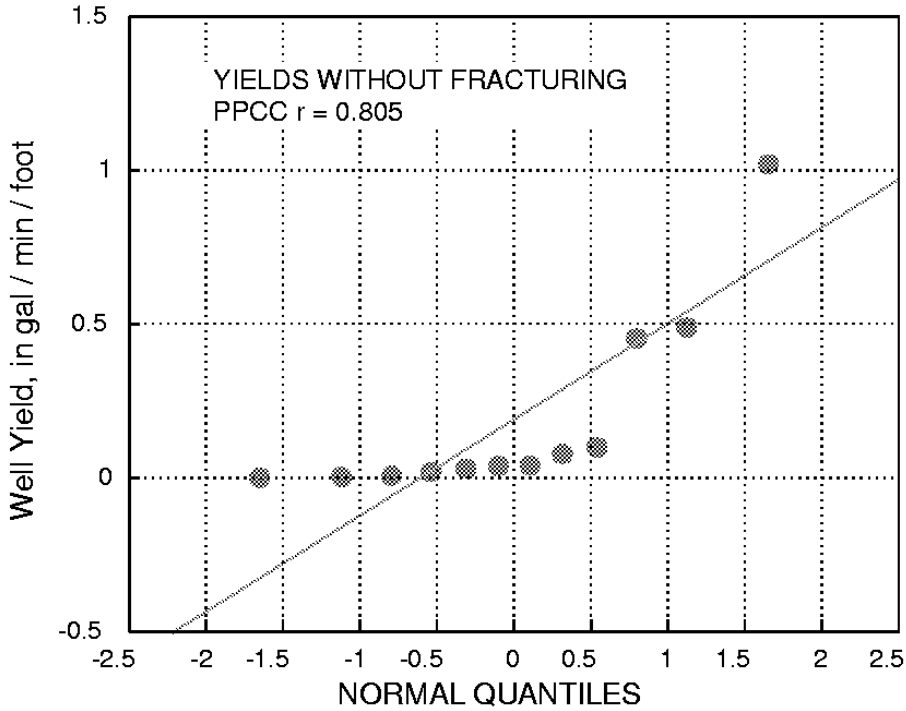


Figure 4.6 Probability plot for the yields without fracturing, with PPCC r

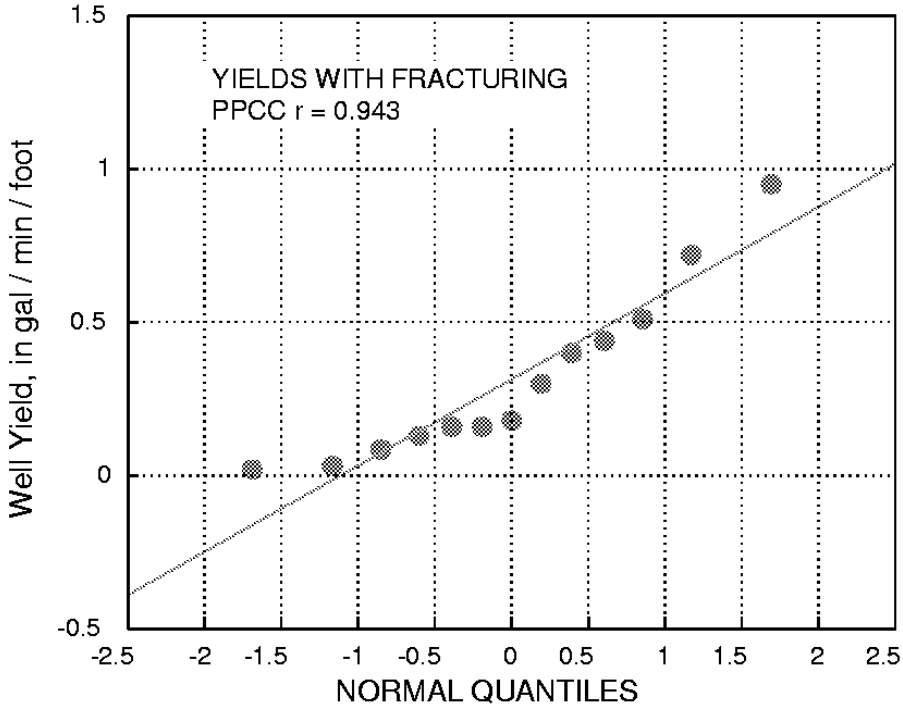


Figure 4.7 Probability plot for the yields with fracturing, with PPCC r

Exercises

4.1 The following are annual streamflows for the Green R. at Munfordville, KY. Beginning in 1969 the stream was regulated by a reservoir.

before				after	
1950	4910	1960	2340	1969	1350
1951	3660	1961	2600	1970	2350
1952	3910	1962	3410	1971	3140
1953	1750	1963	1870	1972	3060
1954	1050	1964	1730	1973	3630
1955	2670	1965	2730	1974	3890
1956	2880	1966	1550	1975	3780
1957	2600	1967	4060	1976	3180
1958	3520	1968	2870	1977	2260
1959	1730			1978	3430
				1979	5290
				1980	2870

Test both before and after data sets for normality using the PPCC test. If either are non-normal, transform the data and re-test in order to find a scale which appears to be close to a normal distribution.

4.2 Test the arsenic data and transformed data of Exercise 2.2 for normality.

Chapter 5

Differences between Two Independent Groups

Wells upgradient and downgradient of a hazardous waste site are sampled to determine whether the concentrations of some toxic organic compound known to reside in drums at the site are greater in the downgradient wells. Are they greater at the $\alpha = 0.01$ significance level? If so, the ground water is declared to be contaminated, and the site will need to be cleaned up.

Measurements of a biological diversity index are made on sixteen streams. Eight of the streams represent "natural" conditions, while the other eight have received urban runoff. Is the biological quality of the urban streams worse than that of the "natural" streams?

Unit well yields are determined for a series of bedrock wells in the Piedmont region. Some wells tap areas where fracturing is prevalent, while other wells are drilled in largely unfractured rock. Does fracturing affect well yields, and if so how?

These are examples of comparisons of two independent groups of data, to determine if one group tends to contain larger values than the other. The data are independent in the sense that there is no natural structure in the order of observations across groups -- there are no pairings of data between observation 1 of group 1 and observation 1 of group 2, etc. Where such a pairing does exist, methods like those of Chapter 6 should be used. In some cases it is known ahead of time which group is expected to be larger (a one-sided test), and in other cases it is not (a two-sided test). This chapter will present and discuss the rank-sum test, a nonparametric procedure for determining whether two independent groups differ. In the special case where the data within each group are known to be normally distributed, and the differences between the groups are additive, the t-test may also be used. Graphical presentations of the test results will be quickly surveyed. Finally, methods for estimating the magnitude of the difference between the two groups are presented, including the Hodges-Lehmann estimator, one of a class of efficient and resistant nonparametric estimators unfamiliar to many water resources scientists.

5.1 The Rank-Sum Test

The rank-sum test goes by many names. It was developed by Wilcoxon (1945), and so is sometimes called the Wilcoxon rank-sum test. It is equivalent to a test developed by Mann and Whitney near the same time period, and the test statistics can be derived one from the other. Thus the Mann-Whitney test is another name for the same test. The combined name of Wilcoxon-Mann-Whitney rank-sum test has also been used.

5.1.1 Null and Alternate Hypotheses

In its most general form, the rank-sum test is a test for whether one group tends to produce larger observations than the second group. It has as its null hypothesis:

H_0: Prob [x > y] = 0.5

where the x are data from one group, and the y are from a second group. In words, this states that the probability of an x value being higher than any given y value is one-half. The alternative hypothesis is one of three statements:

H_1: Prob [x > y] ≠ 0.5 (2-sided test -- x might be larger or smaller than y).

H_2: Prob [x > y] > 0.5 (1-sided test -- x is expected to be larger than y)

H_3: Prob [x > y] < 0.5 (1-sided test-- x is expected to be smaller than y).

Note that no assumptions are made about how the data are distributed in either group. They may be normal, lognormal, exponential, or any other distribution, They may be uni-, bi- or multi-modal. In fact, if the only interest in the data is to determine whether one group tends to produce higher observations, the two groups do not even need to have the same distribution!

Usually however, the test is used for a more specific purpose -- to determine whether the two groups come from the same population (same median and other percentiles), or alternatively whether they differ only in location (central value or median). If both groups of data are from the same population, about half of the time an observation from either group could be expected to be higher than that from the other, so the above null hypothesis applies. However, now it must be assumed that if the alternative hypothesis is true, the two groups differ only in their central value, **though not necessarily in the units being used**. For example, suppose the data are shaped like the two lognormal distributions of figure 5.1. In the original units, the data have different sample medians and interquartile ranges, as shown by the two boxplots. A rank-sum test performed on these data has a p-value of <0.001, leading to the conclusion that they do indeed differ. But is this test invalid because the variability, and therefore the shape, of the two distributions differs? Changing units by taking logs, the boxplots of figure 5.2 result. The logs of the data appear to have different medians, but similar IQR's, and thus the logs of the data appear to differ only in central location. The test statistic and p-value for a rank-sum test computed on these transformed data is **identical** to that for the original units! Nonparametric

tests possess the very useful property of being invariant to power transformations such as those of the ladder of powers. Since only the data **or any power transformation of the data** need be similar except for their central location in order to use the rank-sum test, it is applicable in many situations.

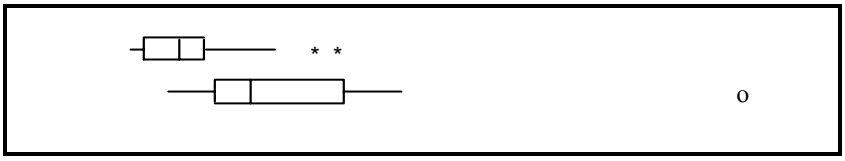


Figure 5.1 Boxplots of two lognormal distributions with different medians and IQRs.

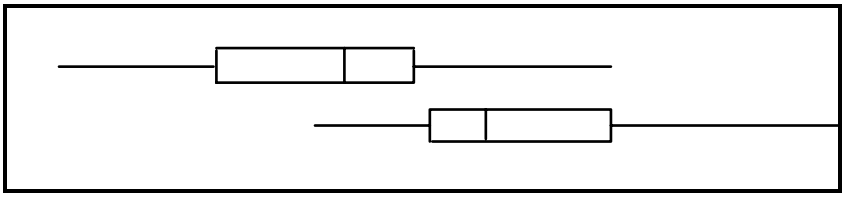


Figure 5.2 Boxplots of the logarithms of the figure 5.1 data. Medians still differ, while IQRs are the same.

5.1.2 Computation of the Exact Test

The exact form of the rank-sum test is given below. It is the only form appropriate for comparing groups of sample size 10 or smaller per group. When both groups have samples sizes greater than 10 (n, m > 10), the large-sample approximation may be used. Remember that computer packages report p-values from the large sample approximation regardless of sample size.

Exact Version of the Rank-Sum test

Situation Two independent groups of data are to be compared. The sample size for the smaller of the two groups x_i, i=1,...n is designated n, while the larger sample size y_j, j=1,...m is designated m.

Test Statistic Compute the joint ranks R_k .
R_k = 1 to (N = n + m), using average ranks in case of ties.

The exact test statistic
W_{rs} = sum of ranks for the group having the smaller sample size,
= ΣR_i i=1,n (use either group when sample sizes are equal: n = m)

Decision Rule. To reject H_0 : Prob [x > y] = 0.5

1. H_1 : Prob [x > y] ≠ 0.5 (the smaller data set tends to have either higher or lower values than the larger data set)
   Reject H_0 if $W_{rs} \leq x^*_{\alpha/2,n,m}$ or $W_{rs} \geq x_{\alpha/2,n,m}$ from Table B4 of the Appendix; otherwise do not reject H_0.

2. H_2 : Prob [x > y] > 0.5 (the smaller data set tends to have higher values than the larger data set)
   Reject H_0 if $W_{rs} \geq x_{\alpha,n,m}$ from Table B4; otherwise do not reject H_0 .

3. H_3 : Prob [x > y] < 0.5 (the smaller data set tends to have lower values than the larger data set)
   Reject H_0 if $W_{rs} \leq x^*_{\alpha,n,m}$ from Table B4; otherwise do not reject H_0 .

Example 1.

Precipitation quality was compared at sites with different land uses by Oltmann and Shulters (1989). A rank-sum test is used to determine if one of the constituents, ammonia plus organic nitrogen, significantly differs (α = 0.05) between the industrial and residential sites.

H_0 : median concentration (industrial) = median concentration (residential)
H_3 : median concentration (industrial) ≠ median concentration (residential).

The 10 observations at each site are assigned ranks from 1 to 20 as follows. Note that three pairs of concentrations (at 0.7, 1.1, and 1.3 mg/L) are tied, and so are assigned tied ranks equal to the average of their two individual ranks:

Ammonia plus organic nitrogen concentration (in mg/L) in precipitation

x_i, y_j = concentrations R_k = joint rank

industrial site				residential site			
x_i	R_k	x_i	R_k	y_j	R_k	y_j	R_k
0.59	4	1.3	14.5	0.3	1	0.9	8
0.87	7	1.6	16	0.36	2	0.92	9
1.1	11.5	1.7	17	0.5	3	1.0	10
1.1	11.5	3.2	18	0.7	5.5	1.3	14.5
1.2	13	4.0	19	0.7	5.5	9.7	20

W_{rs} = sum of the 10 ranks for the residential site (n=m=10, so either could be used)
= 78.5

For this two-sided test, reject H_0 if $W_{rs} \leq x^*_{\alpha/2,n,m}$ or $W_{rs} \geq x_{\alpha/2,n,m}$. From Table B4, $x^*_{.026,10,10} = 79$ and $x^*_{.022,10,10} = 78$. Interpolating halfway between these for $W_{rs} = 78.5$, the p-value for the two-sided test is 0.024•2 = 0.048, and the decision would be to reject H_0 at $\alpha = 0.05$. Reporting the p-value shows how very close the risk of Type I error is to 0.05. The conclusion is therefore that ammonia plus organic nitrogen concentrations from industrial precipitation are significantly different than those in residential precipitation at a p-value of 0.048.

5.1.3 The Large Sample Approximation

For the rank sum test, the distribution of the test statistic W_{rs} closely approximates a normal distribution when the sample size for each group is 10 or above (figure 5.3). With n=m=10, there are 184,756 possible arrangements of the data ranks. The collection of test statistics for each of these comprises the exact distribution of W_{rs}, shown as bars in figure 5.3, with a mean of 105. Superimposed on the exact distribution is the normal distribution which closely approximates the exact values. This demonstrates how well the exact distribution of this test can be approximated, even for relatively small sample sizes. The inset shows a magnified view of the peak of the distribution, with the normal approximation crossing the center of the exact distribution bars.

This approximation does not imply that the data are or must be normally distributed. Rather, it is based on the near normality of the test statistic at large sample sizes. If there are no ties, W_{rs} has a mean μ_W and standard deviation σ_W when H_0 is true of:

$$\mu_W = n \bullet (N+1)/2 \qquad [5.1]$$

$$\sigma_W = \sqrt{n \bullet m \bullet (N+1)/12} \qquad [5.2]$$

where N = n + m.

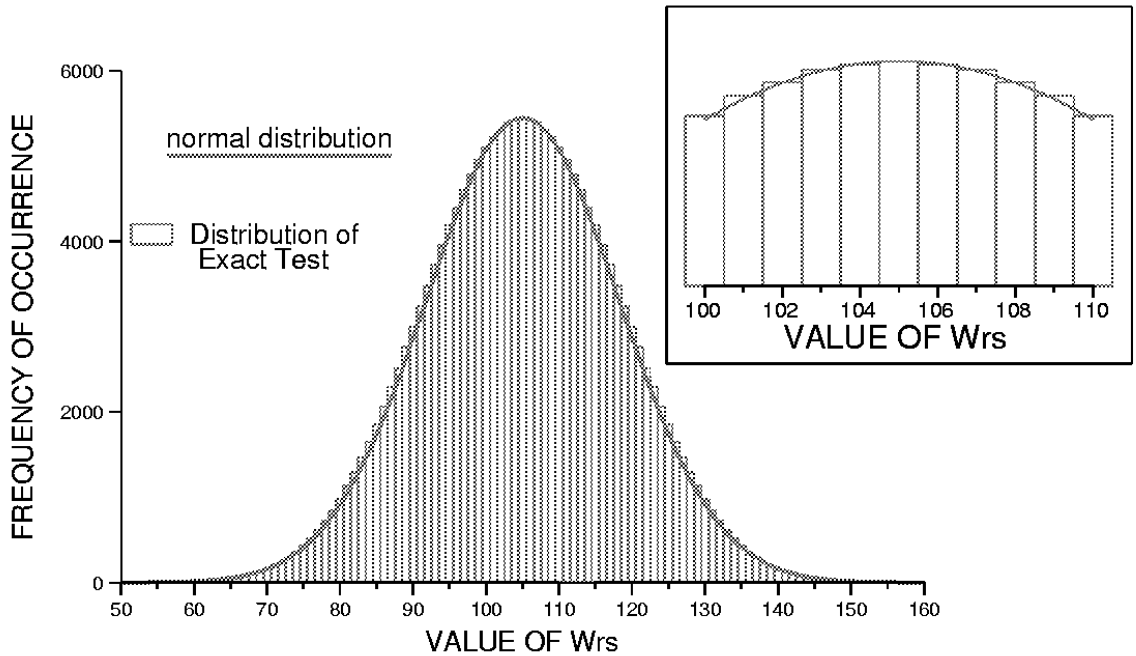


Figure 5.3 Illustration of the distribution of W_{rs} and its fitted normal distribution.

The test statistic for the large sample approximation is computed by standardizing W_{rs} and making a continuity correction. The continuity correction occurs because the normal distribution fits halfway through the top of the bars of the exact test statistic distribution (figure 5.3). The correction moves the probability of occurrence from the outer edge of each bar to its center prior to using the normal curve. It therefore equals d/2, where d is the minimum difference between possible values of the test statistic (the bar width). For the rank-sum test d=1, as the test statistic values change by units of one. Z_{rs}, the standardized form of the test statistic, is therefore computed as

$$Z_{rs} = \begin{cases} \dfrac{W_{rs} - \dfrac{d}{2} - m_W}{s_W} & \text{if } W_{rs} > m_W \\[2ex] 0 & \text{if } W_{rs} = m_W \\[2ex] \dfrac{W_{rs} + \dfrac{d}{2} - m_W}{s_W} & \text{if } W_{rs} < m_W \end{cases} \qquad [5.3]$$

Z_{rs} is compared to a table of the standard normal distribution for evaluation of the test results.

Example 1, cont.

The large-sample approximation is applied to the precipitation nitrogen data. Note that this is inappropriate because there are three pairs of tied values. How close is the approximate to the exact p-value? For the exact test above, $W_{rs} = 78.5$.

$$\mu_W = 10(21)/2 = 105 \qquad \sigma_W = \sqrt{10\ 10\ (21)/12} = 13.23$$

Therefore $Z_{rs} = \dfrac{78.5 + 1/2 - 105}{13.23} = -1.965$

and $p \cong 2 \bullet 0.025 = 0.05$ from a table of the normal distribution such as Table A2 of Iman and Conover (1983). This is very close to the exact test results, and errors decrease with increasing sample sizes.

5.1.3.1 Correction for ties

Conover (1980) presents a further correction to σ_W when ties occur, and tied ranks are assigned. The formula below for σ_{Wt} should be used for computing the large sample approximation rather than σ_W when more than a few ties occur.

$$\sigma_{Wt} = \sqrt{\frac{nm}{N(N-1)}\sum_{k=1}^{N} R_{k^2} - \frac{nm(N+1)^2}{4(N-1)}} \qquad \text{where } N = n+m \qquad [5.4]$$

Example 1, cont.

The tie correction is applied to the large sample approximation for the precipitation nitrogen data. $\sigma_{Wt} = \sqrt{\dfrac{100}{2019}2868.5 - \dfrac{100(21)^2}{419}} = \sqrt{174.61} = 13.21$.

This is essentially identical to the value of 13.23 obtained without the tie correction. The test statistic Z_{rs} and its p-value are unchanged.

5.1.4 The Rank Transform Approximation

Another approximation to the exact rank-sum test is to compute the equivalent parametric test, in this case the t-test, on the ranks R_k rather than on the original data themselves. Computations will be illustrated in detail following the presentation of the t-test in the next section. The rank-transform p-value calculated in that section for the precipitation nitrogen data is 0.042, close to but lower than the exact value, and not as close as the large sample approximation. Rank transform approximations are not as widely accepted as are the large sample approximations. This is due to the fact that the rank transform approximations can result in a lower p-value than the exact test, while the large sample approximation will not. In addition, the rank approximation is often not as close as the large-sample approximation for the

same sample size. Statisticians prefer that an approximation never result in a lower p-value than the exact test, as this means that H_0 will be rejected more frequently than it should. However, this problem only occurs for small sample sizes. For the sample sizes (conservatively, n and m both larger than 25) at which the rank approximation should be used, it should perform well.

5.2 The t-Test

The t-test is perhaps the most widely used method for comparing two independent groups of data. It is familiar to most water resources scientists. However, there are five often overlooked problems with the t-test that make it less applicable for general use than the nonparametric rank-sum test. These are 1) lack of power when applied to non-normal data, 2) dependence on an additive model, 3) lack of applicability for censored data, 4) assumption that the mean is a good measure of central tendency for skewed data, and 5) difficulty in detecting non-normality and inequality of variance for the small sample sizes common to water resources data. These problems were discussed in detail by Helsel and Hirsch (1988), and will be evaluated here in regard to the precipitation nitrogen data.

5.2.1 Assumptions of the Test

The t-test assumes that both groups of data are normally distributed around their respective means, and that they have the same variance. The two groups therefore are assumed to have identical distributions which differ only in their central location (mean). Therefore the t-test is a test for differences in central location only, and assumes that there is an additive difference between the two means, if any difference exists. These are strong assumptions rarely satisfied with water resources data. The null hypothesis is stated as

$H_0 : \mu_x = \mu_y$ the means for groups x and y are identical.

5.2.2 Computation of the t-Test

Two Sample t-test

Situation Two independent groups of data are to be compared. Each group is normally distributed around its respective mean value, and the two groups have the same variance. The sole difference between the groups is that their means may not be the same.

Test Statistic Compute the t-statistic:

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{1/n + 1/m}}$$

where $\bar{x}$ is the sample mean of data in the first group x_i i=1,n
$\bar{y}$ is the sample mean of data in the second group y_j j=1,m
and s is the pooled sample standard deviation, estimating the standard deviation assumed identical in both groups:

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

The sample variances of both groups s_x^2 and s_x^2 are used to estimate s.

Decision Rule. To reject $H_0 : \mu_x = \mu_y$

1. $H_1 : \mu_x \neq \mu_y$ (the two groups have different mean values, but there is no prior knowledge which of x or y might be higher)
   Reject H_0 if $t < -t_{\alpha/2,(n+m-2)}$ or $t > t_{\alpha/2,(n+m-2)}$ from a table of the t distribution; otherwise do not reject H_0.

2. $H_2 : \mu_x > \mu_y$ (prior to seeing any data, x is expected to be greater than y)
   Reject H_0 if $t > t_{\alpha,(n+m-2)}$ from a table of the t distribution; otherwise do not reject H_0 .

3. $H_3 : \mu_x < \mu_y$ (prior to seeing any data, y is expected to be greater than x)
   Reject H_0 if $t < -t_{\alpha,(n+m-2)}$ from a table of the t distribution; otherwise do not reject H_0 .

5.2.3 Modification for Unequal Variances

When the two groups have unequal variances the degrees of freedom and test statistic t should be modified using Satterthwaite's approximation:

Two Sample t-test with Unequal Variances

Situation The mean values of two independent groups of data are to be tested for similarity. Each group is normally distributed around its respective mean value, and the two groups do not have the same variance.

Test Statistic Compute the t-statistic:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/m}}$$

where s_x^2 is the sample variance of the first group, and
s_y^2 is the sample variance of the second group.

Also compute the approximate degrees of freedom df, where

$$df = \frac{(s_x^2/n + s_y^2/m)^2}{\frac{(s_x^2/n)^2}{(n-1)} + \frac{(s_y^2/m)^2}{(m-1)}}$$

Decision Rule. To reject $H_0: \mu_x = \mu_y$

1. $H_1: \mu_x \neq \mu_y$ (the two groups have different mean values, but there is no prior knowledge which of x or y might be higher)
   Reject H_0 if $t < -t_{\alpha/2,(df)}$ or $t > t_{\alpha/2,(df)}$ from a table of the t distribution; otherwise do not reject H_0.

2. $H_2: \mu_x > \mu_y$ (prior to seeing any data, x is expected to be greater than y)
   Reject H_0 if $t > t_{\alpha,(df)}$ from a table of the t distribution; otherwise do not reject H_0 .

3. $H_3: \mu_x < \mu_y$ (prior to seeing any data, y is expected to be greater than x)
   Reject H_0 if $t < -t_{\alpha,(df)}$ from a table of the t distribution; otherwise do not reject H_0 .

Example 1, cont.

The t-test is applied to the precipitation nitrogen data. Are the means of the two groups of data equal? As the variance for the industrial data is 1.2 while for the residential data it is 8.1, Satterthwaite's approximation is used rather than computing an overall variance:

$$t = \frac{1.67 - 1.64}{\sqrt{1.17/10 + 8.12/10}} = 0.03, \text{ and } df = \frac{(1.17/10 + 8.12/10)^2}{\frac{(1.17/10)^2}{9} + \frac{(8.12/10)^2}{9}} = 11.5.$$

Therefore from a table of the t-distribution, the p-value is 0.98. The conclusion: fail to reject H_0. There is essentially no evidence that the means differ using the t-test.

The "t-test on ranks" approximation to the rank-sum test is also computed. This t-test is computed using the joint ranks R_k rather than the original data themselves:

$$t_{rank} = \frac{13.15 - 7.85}{5.4\sqrt{1/10 + 1/10}} = 2.19$$

where 13.15 is the mean rank of the x data, etc. Comparing this to $t_{.025,18} = 2.10$, H_0 is rejected with a p-value of 0.042. The medians are declared different.

5.2.4 Consequences of Violating the t-Test's Assumptions

Computing the probability plot correlation coefficient to test for normality of the two groups of precipitation nitrogen data, the industrial group had a PPCC of 0.895, while the residential group had a PPCC of 0.66. From Table B3 of the Appendix, both correlation coefficients are below the critical value of 0.918 for an α of 0.05, and so both groups must be considered non-normal (see Chapter 4 for details on the PPCC test). A t-test should not have been used on these data. However, if the normality test results are ignored, the t-test declares the group means to be similar, which is commonly interpreted to mean that the two groups are similar. The rank-sum test finds the two groups to be significantly different. This has the following consequences:

1. This example demonstrates the **lack of power** encountered when a t-test is applied to non-normal data. **When parametric tests are applied to non-normal data, their power to detect differences which are truly present is much lower than that for the equivalent nonparametric test** (Bradley, 1968). Thus the t-test is not capable of discerning the difference between the two groups of precipitation nitrogen. The skewness and outliers in the data inflate the sample standard deviation used in the t-test. The t-test assumes it is operating on normal distributions having this standard deviation, rather than on non-normal data with smaller overall spread. It then fails to detect the differences present.

2. As shown by the Q-Q plot of figure 5.5, these data do not exhibit an additive difference between the data sets. A multiplicative model of the differences is more likely, and logs of the data should be used rather than the original units in a t-test. Of course, this is not of concern to the rank-sum test, as the test results will in either units be identical.

3. A t-test cannot be easily applied to censored data, such as data below the detection limit. That is because the mean and standard deviation of such data cannot be computed without either substituting some arbitrary values, or making a further distributional assumption about the

data. This topic is discussed further in Chapter 13. It will only be noted here that all data below a single detection limit can easily be assigned a tied rank, and a rank-sum test computed, without making any distributional assumptions or assigning arbitrary values to the data.

4. The t-test assumes that the mean is a good measure of central tendency for the data being tested. This is certainly not true for skewed data such as the precipitation nitrogen data. The mean of the residential data is greatly inflated by the one large outlier (figure 5.4), making it similar to the mean at the industrial site. The mean is neither resistant to outliers, nor near the center (50th percentile) of skewed data. Therefore tests on the mean often make little sense.

5. When prior tests for normality are used to decide whether a nonparametric test is warranted, departures from normality must be large before they are detected for the small sample sizes ($n<25$ or 30) commonly investigated. In this example, departures were sufficiently drastic that normality was rejected. For lesser departures from normality, computing both the rank sum and t-test would protect against the potential loss of power of the t-test for non-normal data. Alternatively, just the rank sum test could be used for analysis of small data sets.

5.3 Graphical Presentation of Results

In Chapter 2 a detailed discussion of graphical methods for comparisons of two or more groups of data was presented. Overlapping and side-by-side histograms, and dot and line plots of means and standard deviations, inadequately portray the complexities commonly found in water resources data. Probability plots and quantile plots allow complexity to be shown, plotting a point for every observation, but often provide too much detail for a visual summarization of hypothesis test results. Two methods, side-by-side boxplots and Q-Q plots, are very well suited to describing both the results of hypothesis tests, and visually allowing a judgement of whether data fit the assumptions of the test being employed. This is illustrated by the precipitation nitrogen data below.

5.3.1 Side-by-Side Boxplots

The best method for illustrating results of the rank-sum test is side-by-side boxplots. With boxplots only a few quantiles are compared, but the loss of detail is compensated for by greater clarity. In figure 5.4 are boxplots of the precipitation nitrogen data. Note the difference in medians is clearly displayed, as well as the similarity in spread (IQR). The rejection of normality by PPCC tests is seen in the presence of skewness (industrial) and an outlier (residential). Side-by-side boxplots are an effective and concise method for illustrating the basic characteristics of data groups, and of differences between those groups.

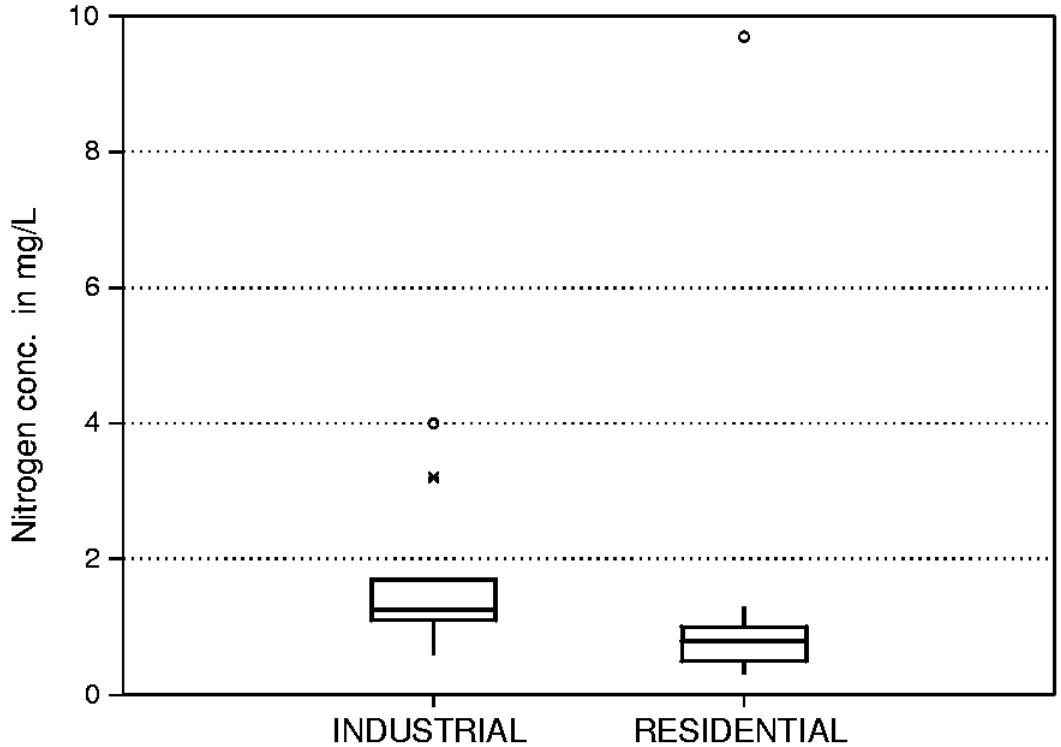


Figure 5.4 Boxplots of the precipitation nitrogen data. Note the skewness and outliers.

5.3.2 Q-Q Plots

Another method for illustration of rank-sum results is the quantile-quantile (Q-Q) plot described in Chapter 2. Quantiles from one group are plotted against quantiles of the second data group. Chapter 2 has shown that when sample sizes of the two groups are identical, the x's and y's can be ranked separately, and the Q-Q plot is simply a scatterplot of the ordered data pairs (x_1 , y_1).....(x_n, y_n). When sample sizes are not equal ($n<m$), the quantiles from the smaller data set are used as is, and the n corresponding quantiles for the larger data set are interpolated.

It is always helpful in a Q-Q plot comparing two groups to plot the y = x line. Figure 5.5 is a Q-Q plot of the precipitation nitrogen data. Two important data characteristics are apparent. First, the data are not parallel to the y = x line, and therefore quantiles do not differ by an additive constant. Instead, they increasingly depart from the line of equality indicating a multiplicative relationship. Note that the Q-Q plot shows that a t-test would not be applicable without a transformation, because it assumes an additive difference between the two groups. The rank-sum test does not make this assumption, and is directly applicable to groups differing by a multiplicative constant (rank procedures will not be affected by a power transformation).

The magnitude of this relationship between two sets of quantiles on a Q-Q plot can be estimated using the median of all possible ratios (y_j/x_i), i=1,n and j=1,n. This is a type of

Hodges-Lehmann estimator, as discussed in the next section. The median ratio equals 0.58, and the line residential = 0.58•industrial is drawn in figure 5.5. Note the resistance of the median ratio to the one large outlier.

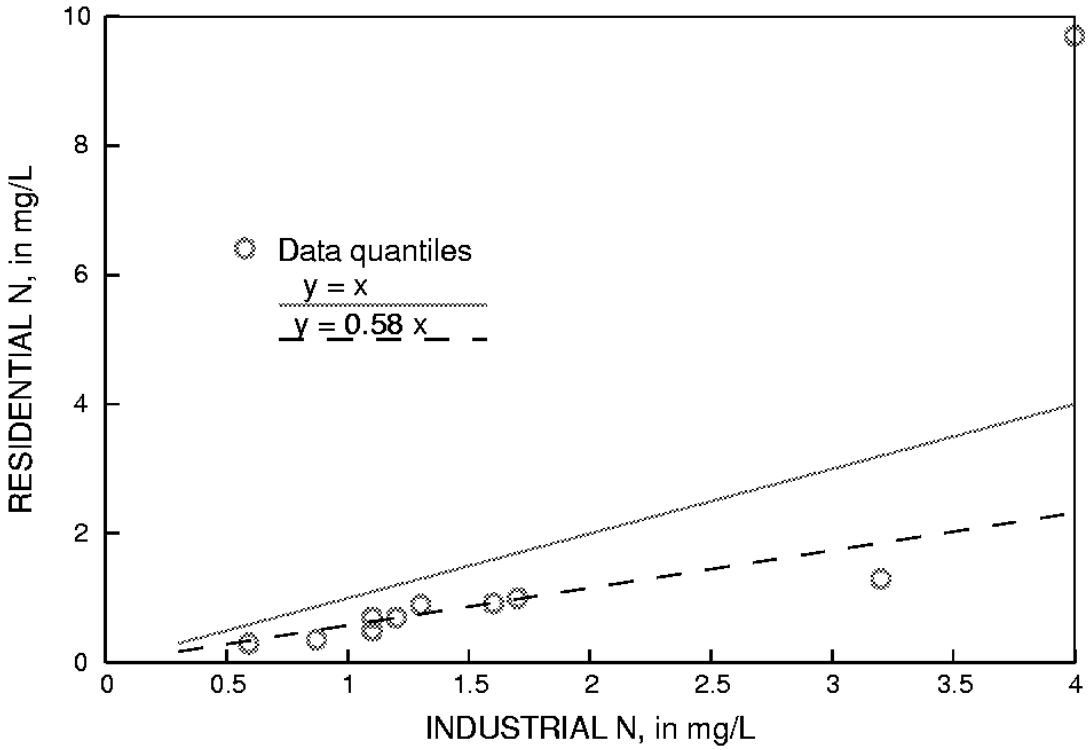


Figure 5.5 Q-Q plot of the precipitation nitrogen data.

Second, the data are crowded together at low concentrations while spread further apart at higher concentrations -- a pattern indicating right-skewness. To remedy both skewness and non-additivity, a power transformation with $\theta < 1$ was chosen, the base 10 log transform ($\theta = 0$). A Q-Q plot of data logarithms is shown in figure 5.6. Note that the data are now spread more evenly from low to high concentrations, indicating skewness has decreased. The slope of the quantiles is now parallel to the y = x line. Thus a multiplicative relationship in original units has become an additive relationship in log units, with the Hodges-Lehmann estimate (see next section) of the difference between log(x) and log(y) $\hat{\Delta}$ equal to −0.237. Note that $\hat{\Delta}$ is the log of the Hodges-Lehmann estimate of the ratios in the original units, $\log_{10}(0.58) = -0.237$. The line parallel to y=x, log(residential) = −0.237•log(industrial), is plotted on figure 5.6. A t-test would now be appropriate for the logarithms, assuming each group's transformed data were approximately normal.

In summary, Q-Q plots of the quantiles of two data groups illustrate the reasonableness of hypothesis tests (t-test or rank-sum), while providing additional insight that the test procedures do not provide. Q-Q plots can demonstrate skewness, the presence of outliers, and inequality of

variance to the data analyst. Perhaps most importantly, the presence of either an additive or multiplicative relationship between the two groups can easily be discerned. Since the t-test requires an additive difference between two groups, Q-Q plots can signal when transformations to produce additivity are necessary prior to using the t-test.

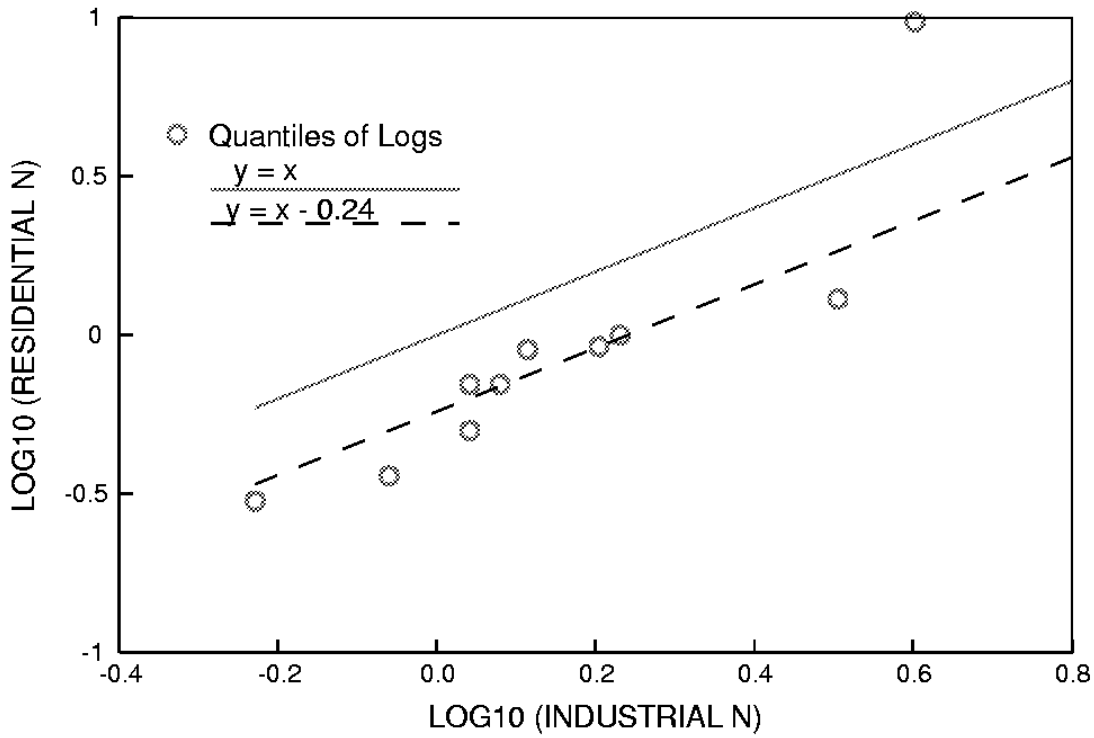


Figure 5.6 Q-Q plot of the logs of the precipitation nitrogen data.

5.4 Estimating the Magnitude of Differences Between Two Groups

After completion of an hypothesis test comparing two groups of data, the logical next step is to determine by how much the two groups differ. The most well-known approach, related to the two-sample t-test, is to compute the difference between the two group means ($\bar{x}-\bar{y}$). A more robust alternative, related to the rank-sum test, is one of a class of nonparametric estimators known as Hodges-Lehmann estimators. These two estimators are compared in the following sections.

5.4.1 The Hodges-Lehmann Estimator

One nonparametric estimate of the difference between two independent groups is a Hodges-Lehmann estimator $\hat{\Delta}$ (Hodges and Lehmann, 1963; Hollander and Wolfe, 1973, p. 75-77). This estimator is the median of all possible pairwise differences between the x values and y values

$$\hat{\Delta} = \text{median}\,[x_i - y_j] \qquad \text{for } x_i,\ i=1,\ldots n \text{ and } y_j,\ j=1,..m. \qquad [5.5]$$

There will be n•m pairwise differences.

Example 2

For the following x's and y's , compute 15 – 8 = 7, 15 – 27 = –12, etc:

x_i	y_j	All Possible Differences $(x_i - y_j)$		
15	8	7	9	17
17	27	–12	–10	–2
25	3	12	14	22
	5	10	12	20

Ranked in order from smallest to largest, the 3•4 = 12 pairwise differences are

–12, –10, –2, 7, 9, 10, 12, 12, 14, 17, 20, 22.

The median of these is the average of the 6th and 7th smallest values, or $\hat{\Delta}$ = 11. Note that the unusual y value of 27 could have been any number greater than 14 and the estimator $\hat{\Delta}$ would be unchanged. Thus $\hat{\Delta}$ is resistant.

The $\hat{\Delta}$ estimator is related to the rank-sum test, in that if $\hat{\Delta}$ were subtracted from each of the x observations, the rank-sum statistic W_{rs} would provide no evidence for rejection of the null hypothesis. In other words, a shift of size $\hat{\Delta}$ makes the data appear devoid of any evidence of difference between x and y when viewed by the rank-sum test.

$\hat{\Delta}$ is a median unbiased estimator of the difference in the medians of populations x and y. That is, the probability of underestimating or overestimating the difference between the median of x and the median of y is exactly one-half. If the populations were both normal, it would be a slightly less efficient estimator of differences in medians (or means) than would the parametric estimator $\bar{x} - \bar{y}$. However, when one or both populations is substantially non-normal, it is a more efficient (lower variance) estimator of this difference.

There is another logical nonparametric estimator of the difference in population medians -- the difference between the sample medians $(x_{med} - y_{med})$. For example 2, $(x_{med} - y_{med})$ = 10.5. Note that the difference in sample medians is not necessarily equal to the median of the differences $\hat{\Delta}$. In addition, $(x_{med} - y_{med})$ is always somewhat more variable (less efficient) than is $\hat{\Delta}$ and so is less desirable.

A modified version of the $\hat{\Delta}$ statistic is used as the estimate of the magnitude of the step trend in the seasonal rank-sum test procedure described by Crawford, Slack, and Hirsch (1983, p. 74).

5.4.2 Confidence Interval for $\hat{\Delta}$

A nonparametric interval estimate for $\hat{\Delta}$ illustrates how variable the difference between the medians might be. No distribution is assumed for the pairwise differences. The interval is

computed by a process similar to that for the confidence interval on the median described earlier. The tabled distribution of the test statistic is entered to find upper and lower critical values at one-half the desired alpha level. These critical values are transformed into ranks. After ordering the n•m pairwise differences from smallest to largest, the differences corresponding to those ranks are the ends of the confidence interval.

For small sample sizes, table B4 for the rank-sum test is entered to find the critical value x* having a p-value nearest to $\alpha/2$. This critical value is then used to compute the ranks R_u and R_l corresponding to the pairwise differences at the upper and lower confidence limits for $\hat{\Delta}$. These limits are the R_lth ranked data points going in from either end of the sorted list of N=n•m pairwise differences.

$$R_l = x^* - \frac{n \bullet (n+1)}{2} \qquad [5.6]$$

$$R_u = N - R_l + 1 \qquad \text{for } N = n \bullet m \qquad [5.7]$$

<u>Example 2, cont.</u>
The N=12 possible pairwise differences between x and y were:

−12, −10, −2, 7, 9, 10, 12, 12, 14, 17, 20, 22.

The median of these ($\hat{\Delta}$) was 11. To determine an $\alpha \cong 0.10$ confidence interval for $\hat{\Delta}$, the tabled critical value x* nearest to $\alpha/2 = 0.05$ is 7 (p=0.057). The rank R_l of the pairwise difference at the lower end of the confidence interval is therefore

$$R_l = 7 - \frac{(3 \bullet 4)}{2} = 1 \quad \text{for n=3 and m=4.}$$

R_u, the rank of the pairwise difference at the upper end of the confidence interval is

$$R_u = 12.$$

With such a small data set, the $\alpha = 2 \bullet 0.057 = 0.014$ confidence limit for $\hat{\Delta}$ is the range of the entire data set (the 1st difference in from either end), or

$$-12 \leq \hat{\Delta} \leq 22.$$

When the large-sample approximation to the rank-sum test is used, a critical value $z_{\alpha/2}$ from the table of standard normal quantiles determines the upper and lower ranks of the pairwise differences corresponding to the ends of the confidence interval. Those ranks are

$$R_l = \frac{N - z_{a/2} \bullet \sqrt{\frac{N\,(n+m+1)}{3}}}{2} \qquad [5.8]$$

$$R_u = \frac{N + z_{a/2} \bullet \sqrt{\frac{N\,(n+m+1)}{3}}}{2} + 1 \qquad [5.9]$$

$$= N - R_l + 1$$

Example 1 cont.

For the precipitation nitrogen data there were N = (10)(10) = 100 possible pairwise differences. $\hat{\Delta}$ would be the average of the 50th and 51st ranked differences. For a 95 percent confidence interval on $\hat{\Delta}$, $z_{\alpha/2} = 1.96$ and

$$R_l = \frac{100 - 1.96 \bullet \sqrt{\frac{100\,(10+10+1)}{3}}}{2} = 24.1$$

$$R_u = 100 - 24.1 + 1 = 76.9$$

the 24.1st ranked slope from either end. Rounding to the nearest integer, the 24th and 77th ranked slopes are used as the ends of the $\alpha \cong 0.05$ confidence limit on $\hat{\Delta}$. Note that using the exact formula, from Table B4 the exact α level is determined to be 2•0.026 = 0.052.

5.4.3 Difference Between Mean Values

As noted above, in the situation where the t-test is appropriate, the difference between the means of both groups $\bar{x} - \bar{y}$ is the most efficient estimator of the difference between the two groups of data. Perhaps obvious is that when x and y are transformed prior to performing the t-test, $(\bar{x} - \bar{y})$ does not estimate the difference between group means in their original units. Less obvious is that a re-transformation of $(\bar{x} - \bar{y})$ back to original units also does not estimate the difference between group means, but is closer to a function of group medians. For the log transformation as an example, $\bar{x} - \bar{y}$ retransformed would equal the ratio of geometric means of the two groups. How close such a re-transformation comes to estimating the ratio of group medians depends on how close the data are to being symmetric in their transformed units.

5.4.4 Confidence Interval for $\bar{x} - \bar{y}$

An interval estimate for the difference in means $\bar{x} - \bar{y}$ is also available. It is appropriate in situations where the t-test may be used -- when both data groups closely follow a normal

distribution. When the variances of the two groups are similar and the pooled standard deviation s is used in the test, the confidence interval is

$$CI = \bar{x} - \bar{y} \pm t_{\alpha/2,(n+m-2)} \bullet s\sqrt{1/n + 1/m} \quad [5.10]$$

When the standard deviations of the two groups are dissimilar and cannot be pooled, the confidence interval becomes

$$CI = \bar{x} - \bar{y} \pm t_{\alpha/2,(df)} \bullet \sqrt{s_x^2/n + s_y^2/m} \quad [5.11]$$

where df is the approximate degrees of freedom used in the t-test.

Exercises

5.1 For the precipitation nitrogen data of Example 1, what would W_{rs} have been had the industrial site been used rather than the arbitrary choice of the residential site. What is the effect on the p-value?

5.2 Historical ground-water quality data for a shallow aquifer underlying agricultural land shows the following nitrate concentrations (mg/L):

pre-1970			post-1970		
1	2	4	1	5	14
1	3	5	2	8	15
1	3	5	2	10	18
2	4	10	4	11	23

Given that we wish to test for a change in concentration between the two periods, should this be a one-sided or two-sided test?

5.3 Annual streamflows for the Green R. at Munfordville, KY were listed in Exercise 4.1. Beginning in 1969 the stream was regulated by a reservoir.

a. Construct a Q-Q plot, and indicate whether the flows exhibit an additive or multiplicative relationship, or neither.
b. Does there appear to be a relationship between (after–before) or (after/before) and the magnitude of annual flow itself? If so, explain why this might occur.
a. Test whether flows after the reservoir came onstream are different.

5.4 Consider the following small data set

X: 1.0, 2.0, 3.0, 4.0

Y: 1.5, 2.5, 3.5, 4.5, 5.5, 7.0, 10.0, 20.0, 40.0, 100.0

Using the Table B4, determine the two-sided p value for an additive difference between the X and Y data using the exact rank-sum test. Then compute it using the large-sample approximation. Then compute it using the t-test on ranks. Compute the expected difference $\hat{\Delta}$ between X and Y.

5.5 Unit well yields, in gallons per minute per foot of water-bearing material, were contrasted for wells within valleys containing fracturing versus valleys with no fracturing (Wright, 1985). For the PPCC test for normality, r(with)=0.943 and r(without)=0.805. Perform the appropriate $\alpha = 0.05$ test to discern whether fracturing is associated with higher mean unit well yield

Yields with fracturing		Yields without	
0.95	0.16	1.02	0.040
0.72	0.16	0.49	0.030
0.51	0.13	0.454	0.020
0.44	0.086	0.10	0.007
0.40	0.031	0.077	0.003
0.30	0.020	0.041	0.001
0.18			

5.6 Assume that the unit well yield data are now trace organic analyses from two sampling sites and that all values below 0.050 were reported as "< 0.05." Retest the hypothesis that $H_0 : \mu_x = \mu_y$ versus $H_1 : \mu_x > \mu_y$ using the rank-sum test. By how much does the test statistic change? Are the results altered by presence of a detection limit? Could a t-test be used in this situation?

Chapter 6

Matched-Pair Tests

To determine the effectiveness of an acid solution in developing wells in carbonate rock, yields of twenty wells were measured both before and after treatment of the wells with acid. Factoring out the differences in yield between wells, have the yields changed as a result of using the acid? What is the magnitude of this change?

Annual sediment loads are measured at two sites over a period of twenty-four years. Both drainage basins are of essentially the same size, and have the same basin characteristics. However, logging has occurred in one basin during the period, but not in the other. Can the year to year variation in load (due to differences in precipitation) be compensated for, to determine whether the site containing logging produced generally higher loads than the other?

Two laboratories are compared in a quality assurance program. Each lab is sent one of a pair of 30 samples split into duplicates in the field, to determine if one lab consistently over- or under-estimates the concentrations of the other. If no difference between the labs is seen, their data may be combined prior to interpretation. The differences between labs must be discerned above the sample to sample differences.

As with the tests of Chapter 5, we wish to determine if one group tends to contain larger values than the other. However, now there is a logical pairing of the observations within each group. Further, there may be a great deal of variability between these pairs, as with the year-to-year pairs of sediment data in the second example above. Both basins exhibit low yields in dry years, and higher yields in wet years. This variability between pairs of observations is noise which would obscure the differences between the two groups being compared if the methods of Chapter 5 were used. Instead, pairing is used to block out this noise by performing tests on the differences between data pairs. Two nonparametric tests are presented for determining whether paired observations differ, the sign test and the signed-rank test. Also presented is the paired t-test, the parametric equivalent which may be used when the differences between pairs are known to be normally distributed. After surveying graphical methods to illustrate the test results, estimators for the difference between the two groups are discussed.

For paired observations (x_i,y_i), i=1,2,...n, their differences $D_i = x_i - y_i$ are computed. The tests in this chapter determine whether x_i and y_i are from the same population -- the null hypothesis -- by analyzing the D_i. If there are differences, the null hypothesis is rejected.

When the D_i's have a normal distribution, a paired t-test can be employed. The paired t-test determines whether the mean of the D_i's equals 0. This is equivalent to stating that the mean of the x_i and the y_i are the same. If the D_i's are symmetric, but not necessarily normal, a signed-rank test can be used. The signed-rank test determines whether the median of the the D_i's is equal to 0. The assumption of symmetry made by the signed-rank test is much less restrictive than that of normality, as there are many non-normal distributions which are symmetric. As a result, the signed-rank test is a more generally applicable method than the t-test. If the differences are asymmetric, the sign test may be used. The sign test does not require an assumption of symmetry or normality. It tests a more general hypothesis than comparisons of means or medians -- does x tend to be higher (or lower, or different) than y? The sign test is the most generally applicable of the three methods. It is also appropriate when the magnitude of the paired differences cannot be computed but one observation can be determined to be higher than the other, as when comparing a <1 to a 3. (Analysis of data below the detection limit is discussed in detail in Chapter 13. See also exercises 6.4 and 6.5 at the end of this chapter.)

6.1 The Sign Test

For data pairs (x_i,y_i) i=1,...n, the sign test determines whether x is generally larger (or smaller, or different) than y, without regard to whether that difference is additive. The sign test may be used regardless of the distribution of the differences, and thus is fully nonparametric.

6.1.1 Null and Alternate Hypotheses

The null and alternative hypotheses may be stated as follows:

H_0: Prob [x > y] = 0.5,

versus one of the three possible alternative hypotheses:

H_1: Prob [x > y] ≠ 0.5 (2-sided test -- x might be larger or smaller than y).

H_2: Prob [x > y] > 0.5 (1-sided test -- x is expected to be larger than y)

H_3: Prob [x > y] < 0.5 (1-sided test-- x is expected to be smaller than y).

6.1.2 Computation of the Exact Test

If the null hypothesis is true, about half of the differences D_i will be positive ($x_i > y_i$) and about half negative ($x_i < y_i$). If one of the alternate hypotheses is true instead, more than half of the differences will tend to be either positive or negative.

The exact form of the sign test is given below. It is the form appropriate when comparing 20 or fewer pairs of samples. With larger sample sizes the large-sample approximation may be used.

Remember that computer packages usually report p-values from the large sample approximation regardless of sample size.

Exact form of the sign test

Situation Two paired groups of data are to be compared, to determine if one group tends to produce larger (or different) values than the other group. No assumptions about the distribution of the differences $D_i = x_i - y_i$, $i = 1,...N$ are required. This means that no assumption is made that all pairs are expected to differ by about the same amount. Numerical values for the data are also not necessary, as long as their relative magnitudes may be determined.

Tied data Ignore all tied data pairs (all $D_i = 0$). Reduce the sample size of the test to the number of nonzero differences n.

Computation Delete all $D_i = 0$ ($x_i = y_i$). The test uses the n nonzero differences $n = N - [\text{number of } D_i = 0]$. Assign a + for all $D_i > 0$, and a − for all $D_i < 0$.

Test Statistic S^+ = the number of +'s, the number of times $x_i > y_i$, $i = 1,...n$.

Decision Rule To reject H_0: Prob [x > y] = 0.5,

1. H_1: Prob [x > y] ≠ 0.5 (the x measurement tends to be either larger or smaller than the y measurement).

   Reject H_0 if $S^+ \geq x_{\alpha/2,n}$ or $S^+ \leq x'_{\alpha/2,n}$ from Table B5; otherwise do not reject H_0.
2. H_2: Prob [x > y] > 0.5 (the x measurement tends to be larger than the y measurement).

   Reject H_0 if $S^+ \geq x_{\alpha,n}$ from Table B5; otherwise do not reject H_0 .
3. H_3: Prob [x > y] < 0.5 (the x measurement tends to be smaller than the y measurement).

   Reject H_0 if $S^+ \leq x'_{\alpha,n}$ from Table B5; otherwise do not reject H_0 .

<u>Example 1.</u>

Counts of mayfly nymphs were recorded in 12 small streams at low flow above and below industrial outfalls. The mayfly nymph is an indicator of good water quality. The question to be considered is whether effluents from the outfalls decreased the number of nymphs found on the streambeds of that region. A Type I risk level α of 0.01 is set as acceptable. Figure 6.1a presents a separate boxplot of the counts for the above and below groups. Both groups are positively skewed. There is a great deal of variability within these groups due to the differences from one stream to another, though in general the counts below the outfalls appear to be smaller. A rank-sum test as in Chapter 5 between the the two groups would be inefficient, as it

would not block out the stream to stream variation (no matching of the pair of above and below counts in each stream). Variation in counts among the streams could obscure the difference being tested for. The natural pairing of observations at the same stream can be used to block out the stream to stream variability by computing the above–below differences in counts for each stream (figure 6.1b). A test is then performed on these differences. Note the asymmetry of the paired differences. They do not appear to all be of about the same magnitude.

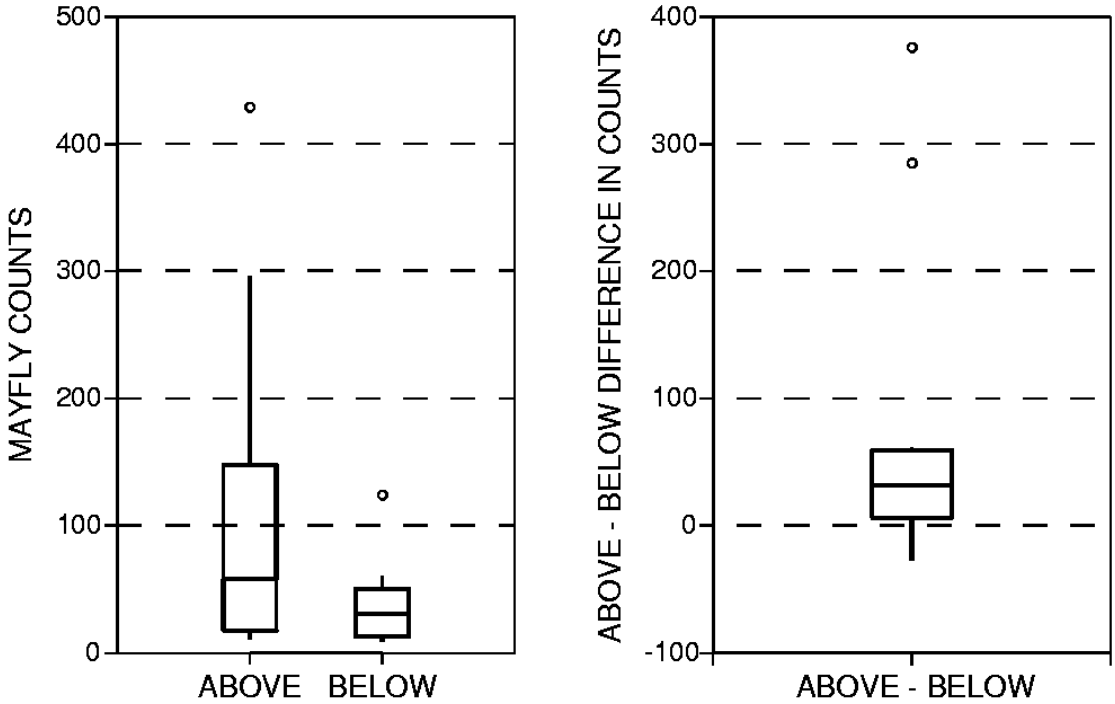


Figure 6.1 a) above and below counts. b) above – below differences.

Table 6.1 Mayfly nymph data.
x_i = counts above outfalls, y_i = counts below outfalls
D_i. = difference $x_i - y_i$.

x_i	y_i	D_i	x_i	y_i	D_i	x_i	y_i	D_i
12	9	3	106	48	58	20	14	6
15	9	6	63	17	46	110	60	50
11	38	−27	296	11	285	429	53	376
41	24	17	53	41	12	185	124	61

The null hypothesis H_0 is that the counts above the outfalls are equally likely to be higher or lower than counts below the outfalls. The one-sided alternate hypothesis H_2 is that the counts below the outfalls are expected to be lower, in which case S^+ would be large.

Of the 12 pairs, 11 are increases, so $S^+ = 11$. Note that this statistic is very resistant to outliers, as the magnitudes of the differences are not used in computing the test statistic. From Table B5 of the Appendix, the one-sided p-value for $S^+ = 11$ is 0.003. Therefore reject that counts above and below the outfall are the same at $\alpha = 0.01$.

6.1.3 The Large Sample Approximation

For sample sizes n>20 the exact sign test statistic can be modified so that its distribution closely follows a standard normal distribution. Again, this does not mean that the data or their differences require normality. It is only the modified test statistic which follows a normal distribution.

The large sample approximation for the sign test takes the standardized form

$$Z^+ = \begin{cases} \dfrac{S^+ - \frac{1}{2} - \mu_{S^+}}{\sigma_{S^+}} & \text{if } S^+ > \mu_{S^+} \\[2ex] 0 & \text{if } S^+ = \mu_{S^+} \\[2ex] \dfrac{S^+ + \frac{1}{2} - \mu_{S^+}}{\sigma_{S^+}} & \text{if } S^+ < \mu_{S^+} \end{cases}$$

$$\boxed{\text{where } \mu_{S^+} = \frac{n}{2}, \text{ and } \sigma_{S^+} = \frac{1}{2}\sqrt{n}\ .}$$

The 1/2 in the numerator of Z^+ is again a continuity correction (see Chapter 5). Z^+ is compared to a table of the standard normal distribution to obtain the approximate p-value. Using the mayfly data of Example 1, the approximate p-value of $p = 0.005$ is obtained below. This is very close to the true (exact) p=0.003, and both are sufficiently small that the decision to reject H_0 would not be altered by their difference.

Therefore, if accurate p-values are of primary concern, such as when p is close to the agreed-upon risk α, and the sample size is 20 or smaller, perform the exact test to get accurate p-values. Regardless of sample size, if p-values are not the primary interest and one must simply decide to reject H_0 or not, when p-values are much smaller (such as 0.001) or much larger (such as 0.50)

than α the decision whether to reject H_0 will be sufficiently clear from the approximate procedure.

<u>Example 1, cont.</u>

For $S^+ = 11$, $\mu_{S^+} = \frac{12}{2} = 6$ $\sigma_{S^+} = \frac{1}{2}\sqrt{12} = 1.73$

$$Z^+ = \frac{11 - \frac{1}{2} - 6}{1.73} = 2.60$$

And from a table of the normal distribution, the approximate one-sided p-value = 0.005.

6.2 The Signed-Rank Test

The signed-rank test was developed by Wilcoxon (1945), and is sometimes called the Wilcoxon signed-rank test. It is used to determine whether the median difference between paired observations equals zero. It may also be used to test whether the median of a single data set is significantly different from zero.

6.2.1 Null and Alternate Hypotheses

For $D_i = x_i - y_i$, the null hypothesis for the signed-rank test is stated as:

H_0: median[D] = 0 .

The alternative hypothesis is one of three statements:

H_1:	median[D] $\neq$ 0	(2-sided test -- x might be larger or smaller than y).
H_2:	median[D] > 0	(1-sided test -- x is expected to be larger than y)
H_3:	median[D] < 0	(1-sided test-- x is expected to be smaller than y).

The signed-rank test is usually stated as a determination of whether the x's and y's come from the same population (same median and other percentiles), or alternatively that they differ only in location (central value or median). If both groups are from the same population, regardless of the shape, about half of the time their difference will be above 0, and half below 0. In addition, the distribution of data above 0 will on average mirror that below 0, so that given a sufficient sample size the differences will be symmetric. They may not be anything like a normal distribution, however. If the alternative hypothesis is true, the differences will be symmetric when x and y come from the same shaped distribution (whatever the shape), differing only in central value (median). This is called an **additive difference** between the two groups, meaning that the variability and skewness within each group is the same for both. Boxplots for the two groups would look very similar, with the only difference being an offset of one from the other. The signed-rank test determines whether this "offset", the magnitude of difference between paired observations, is significantly different from zero. For additive differences (the assumption of symmetric differences is valid), the signed-rank test has more power to detect differences than does the sign test.

In addition, the signed-rank test is also appropriate when the differences are not symmetric in the units being used, but **a logarithmic transformation of both data sets will produce differences which are symmetric**. In such a situation a multiplicative relationship is made into an additive relationship in the logarithms. For example, figure 6.2 displays the differences between two positively skewed distributions. A multiplicative relationship between x and y is suspected, ie. $x = c \bullet y$, where c is some constant. This is a common occurrence with water resources data; data sets having higher median values also often have higher variances than "background" sites with low median values. In the original units the D_i from such data are asymmetric. Changing units by taking the logarithms of the data prior to calculating differences, the boxplot of figure 6.3 results. The log transformation ($\theta = 0$) changes a multiplicative relationship to an additive one: $\log x = \log c + \log y$. The variances of the logs are often made similar by the transformation, so that the logs differ only in central value. The Dl_i, the differences in log units, are therefore much more symmetric than the differences in the original units. The median difference in the logs can then be re-transformed to estimate the median ratio of the original units, $\hat{c}$ = median [y/x] = exp (median [Dl]).

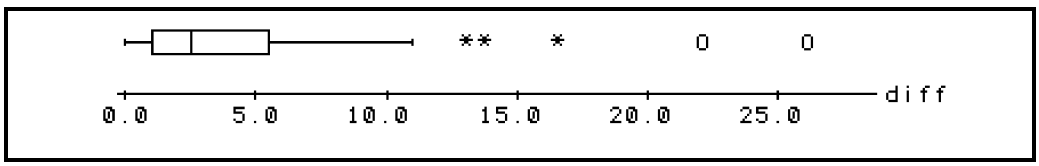


Figure 6.2 Boxplot of asymmetric $D_i = x_i - y_i$

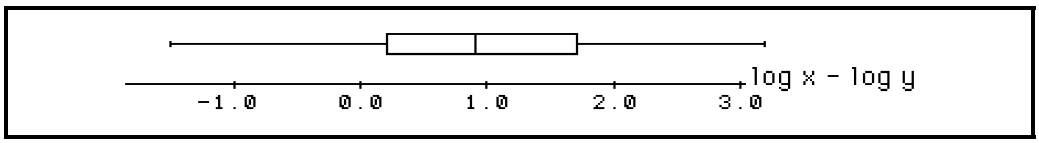


Figure 6.3 Boxplot of symmetric $Dl_i = \log(x_i) - \log(y_i)$

6.2.2 Computation of the Exact Test

If the null hypothesis is true, the median [D] will be close to zero, and the differences will be symmetric around zero. If one of the alternate hypotheses is true instead, the differences will not have a median near zero, but show a symmetric distribution around a nonzero median. Therefore more than half will be either positive or negative. The signed-rank test uses both the signs of the differences as in the sign test, along with the ranks of the absolute values of those differences. This latter information makes sense to use only when the differences are symmetric.

The exact form of the signed-rank test is given below. It is the only form appropriate for comparing 15 or less pairs of samples. With larger sample sizes either large-sample or rank transform approximations may be used.

Exact form of the Wilcoxon signed-ranks test

Situation Two paired groups of data are to be compared, to determine if their differences $D_i = x_i - y_i$ are significantly different from zero. The D_i are assumed to be symmetric. This implies that the two groups differ only in central location.

Computation Compute the absolute value of the differences $|D_i|$, i = 1...N. Rank the $|D_i|$ from smallest to largest. The test uses only nonzero differences, so sample size n= N−[number of D_i=0]. Compute the signed rank R_i, i = 1,...n

$$R_i = \text{rank of } |D_i| \quad \text{for } D_i > 0, \text{ and}$$
$$= -(\text{rank of } |D_i|) \quad \text{for } D_i < 0.$$

Tied data If D_i= 0, delete. When two nonzero differences D_i's are tied, assign the average of the ranks involved to all tied values.

Test Statistic The exact test statistic W^+ is then the sum of all signed ranks R_i having a positive sign:

$$W^+ = \sum_{i=1}^{n} (R_i \& R_i > 0) \qquad \text{where } | \text{ signifies "given that".}$$

Decision Rule To reject H_0: median[D] = 0

1. H_1: median[D] ≠ 0 (the x measurement tends to be either larger or smaller than the y measurement).

   Reject H_0 if $W^+ \geq x_{\alpha/2,n}$ or $W^+ \leq x'_{\alpha/2,n}$ from Table B6; otherwise do not reject H_0.

2. H_2: median[D] > 0 (the x measurement tends to be larger than the y measurement).

   Reject H_0 if $W^+ \geq x_{\alpha,n}$ from Table B6; otherwise do not reject H_0 .

3. H_3: median[D] < 0 (the x measurement tends to be smaller than the y measurement).

   Reject H_0 if $W^+ \leq x'_{\alpha,n}$ from Table B6; otherwise do not reject H_0 .

Example 1, cont.

The differences D_i result in the signed-ranks R_i of table 6.2. From these

$$W^+ = \text{the sum of the positive } R_i\text{'s} = 72.$$

From Table B6, the one-sided p-value for n=12 and $W^+ = 72$ is 0.003. This is strong evidence against the null hypothesis being true. However, the D_i are asymmetric, violating one of the test's assumptions, and indicating that the differences between the two groups may not be an additive one. Asymmetry can be expected to occur when large values tend to produce large differences, and smaller values smaller differences. This indicates that a multiplicative relationship between the data pairs is more realistic. So projecting that a multiplicative relationship may have produced the skewed distribution of D_i's, the base 10 logs of the data were calculated, and a new set of differences

$$Dl_i = \log(x_i) - \log(y_i)$$

are computed and presented in table 6.2 and figure 6.4. Comparing figures 6.4 and 6.1b, note that these Dl_i are much more symmetric than those in the original units. Using the Dl_i,

$$W^+ = \text{the sum of the positive } Rl_i\text{'s} = 69$$

and the exact p-value from Table B6 is 0.008. This should be considered more correct than the results for the untransformed data, as the differences are more symmetric, meeting the requirements of the test procedure. Note that the p-values are not drastically changed, however, and the conclusion to reject H_0 was not affected by the lack of a transformation.

Table 6.2 Mayfly nymph data.

D_i = difference $x_i - y_i$ R_i = signed ranks of D_i

Dl_i = difference of logs Rl_i = signed ranks of Dl_i

D_i	R_i	Dl_i	Rl_i	D_i	R_i	Dl_i	Rl_i	D_i	R_i	Dl_i	Rl_i
3	1	0.125	2	58	9	0.344	8	6	2.5	0.155	3
6	2.5	0.222	5	46	7	0.569	10	50	8	0.263	7
–27	–6	–0.538	–9	285	11	1.430	12	376	12	0.908	11
17	5	0.233	6	12	4	0.111	1	61	10	0.174	4

6.2.3 The Large Sample Approximation

To avoid requiring a large table of exact signed-rank test statistics for all possible sample sizes, the exact test statistic is standardized by subtracting its mean and dividing by its standard deviation so that its distribution closely follows a standard normal distribution. This approximation is valid for sample sizes of n>15.

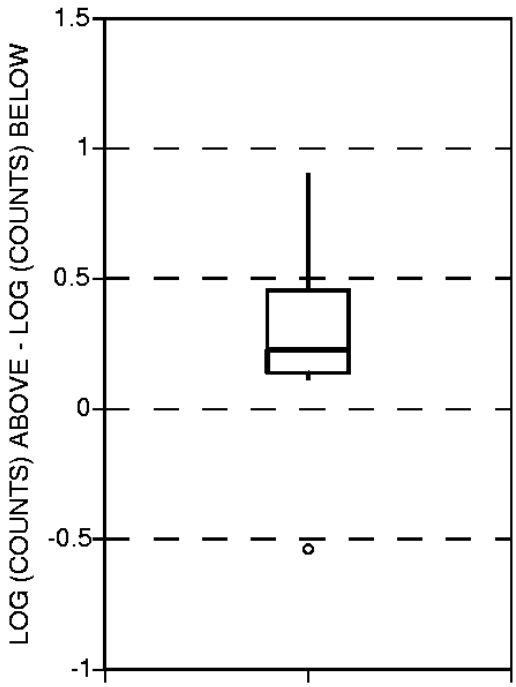


Figure 6.4 Boxplot for the differences of the base 10 logarithms of the mayfly data.

The large sample approximation for the signed-ranks test takes the standardized form

$$Z_{sr}^{+} = \begin{cases} \dfrac{W^{+} - \frac{1}{2} - \mu_{W^{+}}}{\sigma_{W^{+}}} & \text{if } W^{+} > \mu_{W^{+}} \\[2ex] 0 & \text{if } W^{+} = \mu_{W^{+}} \\[2ex] \dfrac{W^{+} + \frac{1}{2} - \mu_{W^{+}}}{\sigma_{W^{+}}} & \text{if } W^{+} < \mu_{W^{+}} \end{cases}$$

$$\boxed{\text{where } \mu_{W^{+}} = \frac{n \bullet (n+1)}{4}, \text{ and } \sigma_{W^{+}} = \sqrt{\frac{n \bullet (n+1) \bullet (2n+1)}{24}}.}$$

The 1/2 in the numerator of Z_{sr}^{+} is the continuity correction. Z_{sr}^{+} is compared to a table of the standard normal distribution to obtain the approximate p-value for the signed-rank test. For the logarithms of the mayfly data of Example 1, the approximate p-value of $p = 0.01$ is obtained below. This is close to the exact value of 0.008, considering that the sample size of 12 is too

small for use of the approximation. When the sample size is 15 or smaller, perform the exact test to get accurate p-values.

<u>Example 1, cont.</u>

For $W^+ = 69$, $\mu_{W^+} = \frac{12 \bullet (13)}{4} = 39$ $\sigma_{W^+} = \sqrt{\frac{12 \bullet (13) \bullet (25)}{24}} = 12.75$

$$Z_{sr}^+ = \frac{69 - \frac{1}{2} - 39}{12.75} = 2.31$$

And from a table of the normal distribution, the approximate one-sided p-value = 0.010.

6.2.4 The Rank Transform Approximation

The rank transform approximation for the signed-rank test is computed by performing a paired t-test on the signed ranks R_i (or Rl_i, if the differences of the logs are more symmetric) rather than on the original data. For this approximation the zero differences $D_i = 0$ are retained prior to computing the test so that there are N, not n, signed ranks. This approximation should be called a "t-test on signed ranks" rather than a signed-ranks test for the sake of clarity. Computations will be given in detail following the presentation of the paired t-test in the next section. The rank-transform p-value calculated in that section for the logs of the mayfly data is 0.005, close to the exact p-value of 0.008. The rank transform approximation should be acceptable for sample sizes greater than 15.

6.3 The Paired t-Test

The paired t-test is the most commonly used test for evaluating matched pairs of data. However, it should not be used without expecting the paired differences D_i to follow a normal distribution. Only if the D_i are normal should the t-test be used. As with the signed-ranks test, logarithms may be taken prior to testing for normality if a multiplicative relationship is suspected. In contrast, all symmetric data, or data which would be symmetric after taking logarithms, may be tested using the signed-ranks test regardless of whether they follow a normal distribution.

6.3.1 Assumptions of the Test

The paired t-test assumes that the paired differences D_i are normally distributed around their mean. The two groups of data are assumed to have the same variance and shape. Thus if they differ, it is only in their mean (central value). The null hypothesis can be stated as

$H_0 : \mu_x = \mu_y$ the means for groups x and y are identical, or

$H_0 : \mu_{[D]} = 0$ the mean difference between groups x and y equals 0.

When the D_i are not normal, and especially when they are not symmetric, the p-values obtained from the t-test will not be accurate. When the D_i are asymmetric, the mean will not provide a

good estimate of the center, as discussed in Chapter 1. Therefore $\mu_{[D]}$ will not be a good estimate of the additive difference between x and y.

6.3.2 Computation of the Paired t-Test

Paired t-test

Situation Two paired groups of data are to be compared, to determine if their differences $D_i = x_i - y_i$ are significantly different from zero. These differences must be normally distributed. Both x and y follow the same distribution (same variance), except that μ_x and μ_y might not be equal.

Test Statistic Compute the paired t-statistic: $t_p = \dfrac{\overline{D}\sqrt{n}}{s}$

where $\overline{D}$ is the sample mean of the differences D_i $\overline{D} = \dfrac{\sum_{i=1}^{n} D_i}{n}$,

and $s = \sqrt{\dfrac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}}$, the sample standard deviation of the D_i's.

Decision Rule. To reject $H_0: \mu_x = \mu_y$

1. $H_1: \mu_x \neq \mu_y$ (the two groups have different mean values, but there is no prior knowledge which of x or y might be higher)
   Reject H_0 if $t_p < -t_{(1-\alpha/2),(n-1)}$ or $t_p > t_{(1-\alpha/2),(n-1)}$ from a table of the t distribution; otherwise do not reject H_0.

2. $H_2: \mu_x > \mu_y$ (prior to seeing any data, x is expected to be greater than y)
   Reject H_0 if $t_p > t_{(1-\alpha),(n-1)}$ from a table of the t distribution; otherwise do not reject H_0 .

3. $H_3: \mu_x < \mu_y$ (prior to seeing any data, y is expected to be greater than x)
   Reject H_0 if $t_p < -t_{(1-\alpha),(n-1)}$ from a table of the t distribution; otherwise do not reject H_0 .

<u>Example 1, cont.</u>

Paired t-test on the mayfly data: The PPCC test for normality on the paired differences D_i has r = 0.82, with an associated p-value of <0.005. Therefore it is highly unlikely that these data come from a normal distribution, and the t-test cannot validly be run. In an attempt to obtain a

distribution closer to normal, the logs of the data are computed. Again as with the signed-rank test, this implies that a multiplicative rather than an additive relationship exists between x and y. The PPCC test for normality of the differences between the logarithms Dl_i has r = 0.92, and a p-value of 0.036. Therefore normality of the logarithms would still be rejected at $\alpha = 0.05$, and the t-test should still not be performed. One could try a series of power transformations, selecting the one whose PPCC test coefficient is closest to 1.0. Howver, it may be difficult to translate the results back into original units -- "the negative square root of differences is statistically different". If the t-test performed on the logs, the following would result:

$$\overline{Dl} = 0.333, \qquad s = \sqrt{\frac{1.59^2}{11}} = 0.479, \quad \text{so } t_p = 2.41.$$

Reject H_0 in favor of H_2 if $t_p > t_{0.95,\,11} = 1.80$. Therefore reject that $\mu_x = \mu_y$. The one-sided p-value for t_p is about 0.02. Note that this is higher than the signed-rank test's p-value of 0.008, reflecting a probable slight loss in power for the t-test as computed on the (non-normal) logarithms of the data.

Rank approximation to the signed-rank test (t-test on signed-ranks): The t-test is performed on the signed-ranks of Dl_i, (see Table 6.2).

$$\overline{Rl} = 5, \qquad s = \sqrt{\frac{18.71^2}{11}} = 5.64, \quad \text{and} \quad t_r = 3.07.$$

Reject H_0 in favor of H_2 if $t_r > t_{0.95,\,11} = 1.80$. Therefore reject H_0. The one-sided p-value equals 0.005, close to the exact p-value of 0.008. Note that the t-test on signed-ranks, as a nonparametric test, ably overlooks the non-normality of the data. The paired t-test does not, and is less able to distinguish the differences between the data logarithms (as shown by its higher p-value) because those differences are non-normal.

6.4 Consequences of Violating Test Assumptions

6.4.1 Assumption of Normality (t-Test)

The primary consequence of overlooking the normality assumption underlying the t-test is a loss of power to detect differences which may truly be present. The second consequence is an unfounded assumption that the mean difference is a meaningful description of the differences between the two groups.

For example, suppose a t-test was blindly conducted on the mayfly data without checking for normality of the differences. The test statistic of t=2.08 has a one-sided p-value of 0.03. This is one order of magnitude above the exact p-value for the (nonparametric) sign test of 0.003. Had an α of 0.01 been chosen, the t-test would be unable to reject H_0 while the sign test would easily reject. The non-normality of the differences "confuses" the t-test by inflating the estimate of standard deviation s, and making deviations from a zero difference difficult to discern.

The mean difference $\overline{D}$ of 74.4 counts for the mayfly data is larger than 10 of the 12 paired differences listed in table 6.1. It has little usefulness as a measure of how many more mayfly nymphs are found above outfalls than below. The lack of resistance of the mean to skewness and outliers heavily favors the general use of the median or Hodges-Lehmann estimator. Another drawback to the mean is that when transformations are used prior to computing a t-test, re-transforming the estimate of the mean difference back into the original units does not provide an estimate of the mean difference in the original units.

6.4.2 Assumption of Symmetry (Signed-Rank Test)

When the signed-rank test is performed on asymmetric differences, it rejects H_0 slightly more often than it should. The null hypothesis is essentially that symmetric differences have a median of zero, and asymmetry favors rejection as does a nonzero median. Some authors have in fact stated that it is a test for asymmetry. However, asymmetry must be severe before a substantial influence is felt on the p-value. While only one outlier can disrupt the t-test's ability to detect differences between two groups of matched pairs, most of the negative differences must be smaller in absolute value than are the positive differences before a signed-rank test rejects H_0 due solely to asymmetry. One or two outliers will have little effect on the signed-rank test, as it uses their rank and not their value itself for the computation. Therefore violation of the symmetry assumption of the signed-rank test produces p-values only slightly lower than they should be, while violating the t-test's assumption of normality can produce p-values much larger than what is correct. Add to this the fact that the assumption of symmetry is less restrictive than that of normality, and the signed-rank test is seen to be relatively insensitive to violation of its assumptions as compared to the t-test.

Inaccurate p-values for the signed-rank test is therefore not the primary problem caused by asymmetry. The p-values for the mayfly data, for example, are not that different ($p = 0.003$ for the original units and 0.008 for the logs) before and after a transformation to achieve symmetry. Both are similar to the p-value for the sign test, which does not require symmetry. However, inappropriate estimates of the magnitude of the difference between data pairs will result from estimating an additive difference when the evidence points towards a multiplicative relationship. Therefore symmetry is especially important to check if the magnitude of the difference between data pairs is to be estimated. Equally as important to check is the form of the relationship between x and y, using the scatterplots of the next section.

6.5 Graphical Presentation of Results

Methods for illustrating matched-pair test results are those already given in Chapter 2 for illustrating a single data set, as the differences between matched pairs constitute a single data set. A probability plot of the paired differences, for example, will show whether or not the data follow a normal distribution. Of the methods in Chapter 2, the boxplot is the single graphic

which best illustrates both the test results and the degree of conformity to the test's assumptions. The equivalent graphic to a Q-Q plot for paired data is a scatterplot of the data pairs. The addition of the x=y line and a smooth of the paired data will help illustrate the test results.

6.5.1 Boxplots

The best method for directly illustrating the results of the sign, signed-rank or paired t-tests is a boxplot of the differences, as in figure 6.1b. The number of data above and below zero and the nearness of the median difference to zero are clearly displayed, as is the degree of symmetry of the D_i. Though a boxplot is an effective and concise way to illustrate the characteristics of the differences, it will not show the characteristics of the original data. This can be better done with a scatterplot.

6.5.2 Scatterplots With X=Y Line

Scatterplots illustrate the relationships between the paired data (figure 6.5). Each (x,y) pair is plotted as a point. Similarity between the two groups of data is illustrated by the closeness of the data to the x=y line. If x is generally greater than y, most of the data will fall below the line. When y exceeds x, the data will lie largely above the x=y line. This relationship can be made clearer for large data sets by superimposing a lowess smooth (see Chapter 10) of the paired data onto the plot.

Data points (or their smooth) generally parallel to the x=y line on the scatterplot would indicate an additive difference between the (x,y) data pairs. Therefore the line x = y + d could be plotted on the figure to illustrate the magnitude of the difference between x and y, where d is the appropriate estimate of the difference between x and y as described in the next section. In figure 6.6 the line x = y + 31.5 is plotted, where 31.5 is the median difference. For an additive relationship the data points would scatter around this line. Obviously the differences do not appear to be additive.

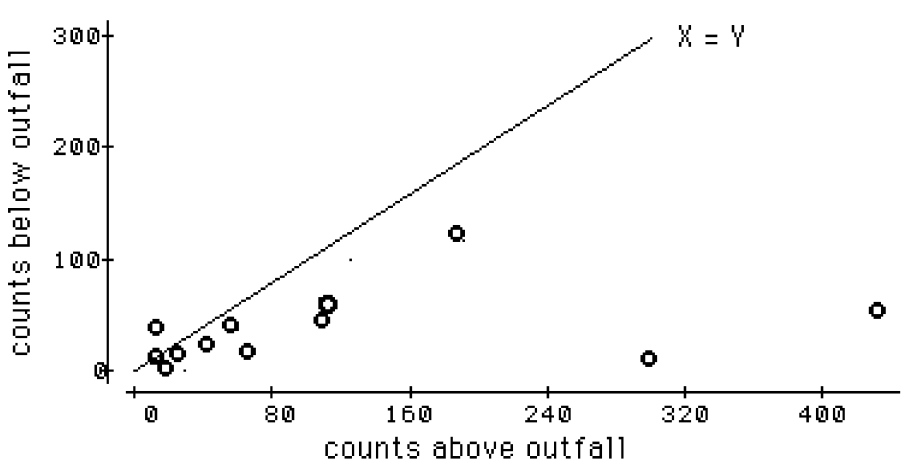


Figure 6.5 Scatterplot of the example 1 mayfly data.

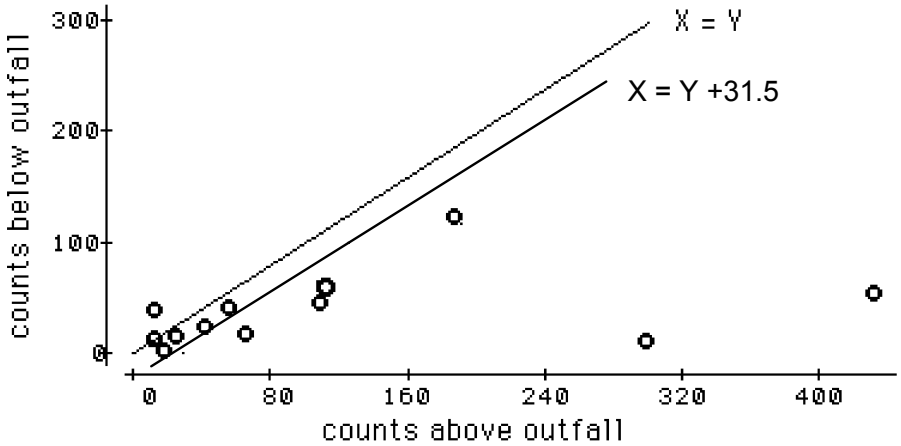


Figure 6.6 Mayfly data with ill-fitting additive relationship x = y+31.5.

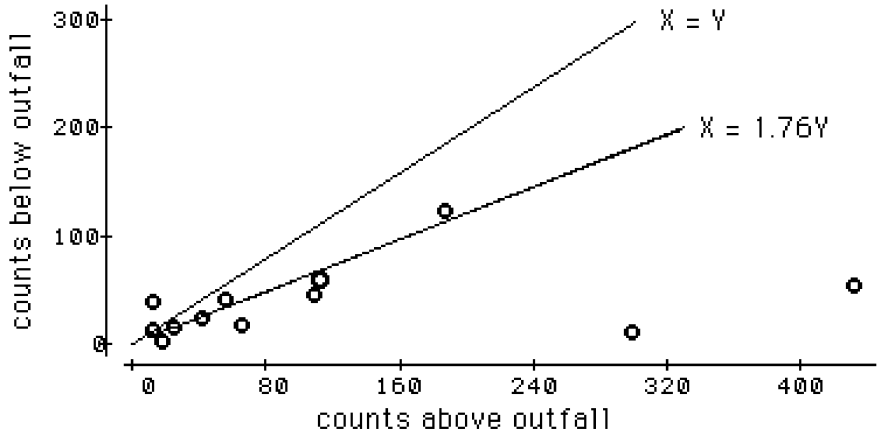


Figure 6.7 Mayfly data with multiplicative relationship x = y•1.76.

Alternatively, an increasing difference between the data and the x=y reference line indicates that there is a multiplicative difference between x and y, requiring a logarithmic transformation prior to the signed-rank or t-test. For a multiplicative relation the line $x = y \bullet f^{-1}(d)$ can be plotted as an aid in visualizing the relation between x and y. For base 10 logs, $f^{-1}(d) = 10^d$ while for natural logs it is exp(d). The mayfly data of example 1 exhibit such a multiplicative relationship, as shown in figure 6.7. There $d = \hat{\Delta}$, the Hodges-Lehmann estimate in log units, resulting in the line x = y•1.76.

6.6 Estimating the Magnitude of Differences Between Two Groups

After testing for differences between matched pairs, a measure of the magnitude of that difference is usually desirable. If outliers are not present, and the differences can be considered normal, an efficient estimator is the mean difference $\overline{D}$. This estimator is appropriate whenever the paired t-test is valid. When outliers or non-normality are suspected, a more robust estimator is necessary. The estimator associated with the signed-rank test is a Hodges-Lehmann estimator $\hat{\Delta}$ (Hollander and Wolfe, 1973) . $\hat{\Delta}$ is the median of all possible pairwise averages of the differences. When the D_i are not symmetric and the sign test is used, the associated estimate of difference is simply the median of the differences D_{med}.

6.6.1 The Median Difference (Sign Test)

For the mayfly data of example 1, the median difference in counts is 31.5. As these data are asymmetric, there is no statement that the two groups are related in an additive fashion. But subtracting this median value from the x data (the sites above the outfalls) would produce data having no evidence for rejection of H_0 as measured by the sign test. Therefore the median is the most appropriate measure of how far from "equality" the two groups are in their original units. Half of the differences are larger, and half smaller, than the median.

A confidence inteval on this difference is simply the confidence interval on the median previously presented in Chapter 4.

6.6.2 The Hodges-Lehmann Estimator (Signed-Rank Test)

Hodges-Lehmann estimators are computed as the median of all possible appropriate combinations of the data. They are associated with many nonparametric test procedures. For the matched-pairs situation, $\hat{\Delta}$ is the median of the n•(n+1)/2 possible pairwise averages:

$$\hat{\Delta} = \text{median } [A_{ij}] \quad \text{where } A_{ij} = [(D_i+D_j)/2] \quad \text{for all } i \le j \qquad [6.1]$$

$\hat{\Delta}$ is related to the signed-rank test in that subtracting $\hat{\Delta}$ from all paired differences (or equivalently, from the x's or y's, whichever is larger) would cause the signed-rank test to have W^+ close to 0, and find no evidence of difference between x and y. For the cases of symmetric differences where the signed-rank test is appropriate, the Hodges-Lehmann estimator more efficiently measures the additive difference between two data groups than does the sample median of the differences D_{med}. For the mayfly data, $\hat{\Delta}$ of the logs = 0.245. The log of upstream counts minus 0.245 estimates the log of the counts below the outfalls. Thus the counts above the outfalls divided by $10^{0.245}$ = 1.76 best estimates the counts below the outfalls (the line X = 1.76 Y in figure 6.7).

6.6.2.1 Confidence interval on $\hat{\Delta}$

A nonparametric interval estimate of the difference between paired observations is computed by a process similar to that for the confidence interval for other Hodges-Lehmann estimators. The tabled distribution of the test statistic is entered to find upper and lower critical values at one-half the desired alpha level. These critical values are transformed into ranks. The pairwise differences A_{ij} are ordered from smallest to largest, and those corresponding to the computed ranks are the ends of the confidence interval.

For small sample sizes, table B6 for the signed-rank test is entered to find the critical value x' having a p-value nearest to $\alpha/2$. This critical value is then used to compute the ranks R_u and R_l corresponding to the pairwise averages A_{ij} at the upper and lower confidence limits for $\hat{\Delta}$. These limits are the R_lth ranked A_{ij} going in from either end of the sorted list of n(n+1)/2 differences.

$$R_l = x' \qquad \text{for } x' = (\alpha/2)\text{th quantile of signed-rank test statistic} \qquad [6.2]$$

$$R_u = x + 1 \qquad \text{for } x = (1-\alpha/2)\text{th quantile of signed-rank test statistic} \qquad [6.3]$$

<u>Example 1, cont.</u>

For the n=12 logarithms of the mayfly data, there are N=78 pairwise averages. For an $\alpha \cong 0.05$ confidence interval, x'=14 and x=64 from table B6 (α =2•0.026 = 0.052). The confidence interval is composed of the 14th and 65th ranked averages (the 14th average in from either end.

For larger sample sizes where the large-sample approximation is used, a critical value $z_{\alpha/2}$ from the table of standard normal quantiles determines the upper and lower ranks of the pairwise averages A_{ij} corresponding to the ends of the confidence interval. Those ranks are

$$R_l = \frac{N - z_{\alpha/2} \bullet \sqrt{\frac{n\,(n+1)\,(2n+1)}{6}}}{2} \qquad [6.4]$$

$$R_u = \frac{N + z_{\alpha/2} \bullet \sqrt{\frac{n\,(n+1)\,(2n+1)}{6}}}{2} + 1 \qquad [6.5]$$

$$= N - R_l + 1 \qquad \text{where } N = n(n+1)/2$$

<u>Example 1 cont.</u>

For the mayfly data with N=78 and n=12, an approximate α=0.05 confidence interval is between the 14th and 65th ranked averages, as computed below:

$$R_l = \frac{78 - 1.96 \bullet \sqrt{\frac{12\ (13)\ (25)}{6}}}{2} = 14.0$$

$$R_u = 78 - 14 + 1 = 65.$$

6.6.3 Mean Difference (t-Test)

For the situation where the differences are not only symmetric but normally distributed and the t-test is used, the most efficient estimator of the difference between the two groups is the mean difference $\overline{D}$. However, $\overline{D}$ is only slightly more efficient than is $\hat{\Delta}$, so that when the data depart from normality even slightly the Hodges-Lehmann estimator is just as efficient as $\overline{D}$. This mirrors the power characteristics of their associated tests, as the signed-rank test is as efficient as the t-test for only slight departures from normality (Lehmann, 1975). Therefore when using "real data" which is never "exactly normal" the mean difference has little advantage over $\hat{\Delta}$, while $\hat{\Delta}$ is more appropriate in a wider number of situations -- for data which are symmetric but not normal.

6.6.3.1 Confidence interval on the mean difference

A confidence interval on the mean difference $\overline{D}$ is computed exactly like any confidence interval for a mean:

$$CI = \overline{D} \pm t_{\alpha/2,(n-1)} \frac{s}{\sqrt{n}} \qquad [6.6]$$

where s is the standard deviation of the differences D_i.

Exercises

6.1 Test the null hypothesis that the median of annual flows for the Conecuh R. at Brantley, Ala. (data in Appendix C2) is 683 cfs for 1941 - 1960. The alternate hypothesis is that it is less than 683 cfs, and alpha = 0.05.

6.2 Which of the following are not matched pairs?

a. analyses of same standard solutions sent to two different laboratories
b. daily sediment discharges above and below a reservoir
c. nitrate analyses from randomly selected wells in each of two aquifers
d. all of the above are matched pairs.

6.3 The following values of specific conductance were measured on the two forks of the Shenandoah River in Virginia (D. Lynch, personal communication).

a. State the appropriate null and alternate hypotheses to see if conductance values are the same in the two forks.
b. Determine whether a parametric or nonparametric test should be used.
c. Compute an $\alpha = .05$ test and report the results.
d. Illustrate and check the results with a plot.
e. Estimate the amount by which the forks differ in conductance, regardless of the test outcome.

Date	South Fork	North Fork	Date	South Fork	North Fork
5-23-83	194	255	2-22-84	194	295
8-16-83	348	353	4-24-84	212	199
10-05-83	383	470	6-04-84	320	410
11-15-83	225	353	7-19-84	340	346
1-10-84	266	353	8-28-84	310	405

6.4 Atrazine concentrations in shallow groundwaters were measured by Junk et al. (1980) before (June) and after (September) the application season. The data are given in Appendix C4. Determine if concentrations of atrazine are higher in groundwaters following surface application than before.

6.5 Try performing the comparison of atrazine concentrations in 6.4 using a t-test, setting all values below the detection limit to zero. Compare the results with those of 6.4. Discuss why the results are similar or different.

Chapter 7

Comparing Several Independent Groups

Concentrations of volatile organic compounds are measured in shallow ground waters across a several county area. The wells sampled can be classified as being contained in one of seven land-use types: undeveloped, agricultural, wetlands, low-density residential, high-density residential, commercial, and industrial/transportation. Do the concentrations of volatiles differ between these types of surface land-use, and if so, how?

Alkalinity, pH, iron concentrations, and biological diversity are measured at low flow for small streams draining areas mined for coal. Each stream drains either unmined land, land strip-mined and then abandoned, or land strip-mined and then reclaimed. The streams also drain one of two rock units, a sandstone or a limestone formation. Do drainages from mined and unmined lands differ in quality? What affect has reclamation had? Are there differences in chemical or biological quality due to rock type separate and distinct from the effects due to mining history?

Three methods for field extraction and concentration of an organic chemical are to be compared at numerous wells. Are there differences among concentrations produced by the extraction processes? These must be discerned above the well-to-well differences in concentration which contribute considerable noise to the data.

The methods of this chapter can be used to answer questions such as those above. These methods are extensions of the ones introduced in Chapters 5 and 6, where now more than two groups of data are to be compared. The classic technique in this situation is analysis of variance. More robust nonparametric techniques are also presented for the frequent situations where data do not meet the assumptions of analysis of variance.

Suppose a set of continuous data, such as concentration or water levels, is collected. It is suspected that one or more influences on the magnitude of these data comes from grouped variables, variables whose values are simply "from group X". Examples include season of the year ("from summer","winter", etc.), aquifer type, land-use type, and similar groups. Each observation will be classified into one of these groups.

First consider the effect of only one grouped variable, calling it an **explanatory variable** because it is believed to explain some of the variation in magnitude of the data at hand. This variable is also called a **factor**. It consists of a set of k groups, with each data point belonging in one of the k groups. For example, the data could be calcium concentrations from wells in one of k aquifers, and the objective is to determine whether the calcium concentrations differ among the k aquifers. Within each group (aquifer) there are n_j observations (the sample size of each group is not necessarily the same). Observation y_{ij} is the ith of n_j observations in group j, so that $i=1,...n_j$ for the jth of k groups $j=1,...k$. The total number of observations N is thus

$$N = \sum_{j=1}^{k} n_j \,, \quad \text{which simplifies to} \qquad N = k \bullet n$$

when the sample size $n_j = n$ for all k groups (equal sample sizes).

The tests in this chapter determine if all k groups have the same central value (median or mean, depending on the test), or whether at least one of the groups differs from the others. When data within each of the groups are normally distributed and possess identical variances, an analysis of variance (ANOVA) can be used. Analysis of variance is a parametric test, determining whether each group's mean is identical. When there are only two groups, the ANOVA becomes identical to a t-test. Thus ANOVA is like a t-test between three or more groups of data, and is restricted by the same types of assumptions as was the t-test. When every group of data cannot be assumed to be normally distributed or have identical variance, a nonparametric test should be used instead. The Kruskal-Wallis test is much like a rank-sum test extended to more than two groups. It compares the medians of groups differentiated by one explanatory variable (one factor).

When the effect of more than one factor is to be evaluated simultaneously, such as both rock type and mining history in one of the examples which began this chapter, the one-way tests can no longer be used. For data which can be assumed normal, several factors can be tested simultaneously using multi-factor analysis of variance. However, the requirements of normality and equal variance now apply to data grouped by each unique combination of factors. This becomes quite restrictive and is rarely met in practice. Therefore nonparametric alternatives are also presented.

The following sections begin with tests for differences due to one factor. Subsequent sections discuss tests for effects due to more than one factor. All of these have as their null hypothesis

that each group median (or mean) is identical, with the alternative that at least one is different. However, when the null hypothesis is rejected, these tests do not tell which group or groups are different! Therefore sections also follow on multiple comparison tests -- tests performed after the ANOVA or Kruskal-Wallis null hypothesis has been rejected, for determining which groups differ from others. A final section on graphical display of results finishes the chapter.

7.1 Tests for Differences Due to One Factor

7.1.1 The Kruskal-Wallis Test

The Kruskal-Wallis test, like other nonparametric tests, may be computed by an exact method used for small sample sizes, by a large-sample approximation (a chi-square approximation) available on statistical packages, and by ranking the data and performing a parametric test on the ranks. Tables for the exact method give p-values which are exactly correct. The other two methods produce approximate p-values that are only valid when sample sizes are large, but do not require special tables. Tables of exact p-values for all sample sizes would be huge, as there are many possible combinations of numbers of groups and sample sizes per group. Fortunately, large sample approximations for all but the smallest sample sizes are very close to their true (exact) values. Thus exact computations are rarely required. All three versions have the same objective, as stated by their null and alternate hypotheses.

7.1.1.1 Null and alternate hypotheses

In its most general form, the Kruskal-Wallis test has the following null and alternate hypotheses:

H_0: All of the k groups of data have identical distributions, versus

H_1: At least one group differs in its distribution.

No assumptions are required about the shape(s) of the distributions. They may be normal, lognormal, or anything else. If the alternate hypothesis is true, they may have different distributional shapes. In this form, the only interest in the data is to determine whether all groups are identical, or whether some tend to produce observations different in value than the others. This difference is not attributed solely to a difference in median, though that is one possibility. Thus the Kruskal-Wallis test, like the rank-sum test, may be used to determine the general equivalence of groups of data.

In practice, the test is usually performed for a more specific purpose -- to determine whether all groups have the same median, or whether at least one median is different. This form requires that all other characteristics of the data distributions, such as spread or skewness, are identical -- though not necessarily in the original units. Any data for which a monotonic transformation, such as in the ladder of powers, produces similar spreads and skewness are also valid. This parallels the rank-sum test (see Chapter 5). As a test for difference in medians, the Kruskal-Wallis null and alternate hypotheses are:

H_0: The medians of the k groups are identical,
H_1: At least one median differs from the others. (a 2-sided test).

As with the rank-sum test, the Kruskal-Wallis test statistic and p-value computed for data that are transformed using any monotonic transformation are identical to the test statiistic and p-value using the original units. Thus there is little incentive to search for transformations (to normality or otherwise) -- the test is applicable in many situations.

7.1.1.2 Computation of the exact test

The exact form of the Kruskal-Wallis test is required when comparing 3 groups with sample sizes of 5 or less per group, or with 4 or more groups of size 4 or less per group (Lehmann, 1975). For larger sample sizes the large-sample approximation is sufficiently accurate. As there are few instances where sample sizes are small enough to warrant using the exact test, exact tables for the Kruskal-Wallis test are not included in this book. Refer to either Conover (1980) or Lehmann (1975) for those tables.

Should the exact test be required, compute the exact test statistic K as shown in the large sample approximation of the following section. K is computed identically for both the exact form or large sample approximation. When ties occur, the large sample approximation must be used.

7.1.1.3 The large-sample approximation

To compute the test, the data are ranked from smallest to largest, from 1 to N. At this point the original values are no longer used; their ranks are used to compute the test statistic. If the null hypothesis is true, the average rank for each group should be similar, and also be close to the overall average rank for all N data. When the alternative hypothesis is true, the average rank for some of the groups will differ from others, reflecting the difference in magnitude of its observations. Some of the average group ranks will then be significantly higher than the overall average rank for all N data, and some will be lower. The test statistic K uses the squares of the differences between the average group ranks and the overall average rank, to determine if groups differ in magnitude. K will equal 0 if all groups have identical average ranks, and will be positive if group ranks are different. The distribution of K when the null hypothesis is true can be approximated quite well by a chi-square distribution with k−1 degrees of freedom.

The degrees of freedom is a measure of the number of independent pieces of information used to construct the test statistic. If all data are divided by the overall group mean to standardize the data set, then when any k−1 average group ranks are known, the final (kth) average rank can be computed from the others as

$$\bar{R}_k = \frac{N}{n_k} \cdot \left(1 - \sum_{j=1}^{k-1} \frac{n_j}{N} \bar{R}_j \right)$$

Therefore there are actually only k−1 independent pieces of information represented by the k average group ranks. From these the kth average rank is fixed.

Large Sample Approximation for the Kruskal-Wallis test

Situation Several groups of data are to be compared, to determine if their medians are significantly different. For a total sample size of N, the overall average rank will equal (N+1)/2. If the average rank within a group (average group rank) differs considerably from this overall average, not all groups can be considered similar.

Computation All N observations are jointly ranked from 1 to N, smallest to largest. These ranks R_{ij} are then used for computation of the test statistic. Within each group, the average group rank $\overline{R}_j$ is computed:

$$\overline{R}_j = \frac{\sum_{i=1}^{n_j} R_{ij}}{n_j} .$$

Tied data When observations are tied, assign the average of their ranks to each.

Test Statistic The average group rank $\overline{R}_j$ is compared to the overall average rank $\overline{R} = (N+1)/2$, squaring and weighting by sample size, to form the test statistic K:

$$K = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left[\overline{R}_j - \frac{N+1}{2} \right]^2 .$$

Decision Rule To reject H_0: all groups have identical distributions, versus H_1: at least one distribution differs

Reject H_0 if $K \geq \chi^2_{1-\alpha,(k-1)}$ the $1-\alpha$ quantile of a chi-square distribution with $(k-1)$ degrees of freedom; otherwise do not reject H_0.

Example 1.

Fecal coliforms, in organisms per 100 ml, were measured in the Illinois River from 1971 to 1976 (Lin and Evans, 1980). A small subset of those data are presented here. Do all four seasons exhibit similar values, or do one or more seasons differ? Boxplots for the four seasons are shown in figure 7.1.

Table 7.1 Selected fecal coliform data (from Lin and Evans, 1980).
[counts in organisms per 100 ml]

	Summer	Fall	Winter	Spring
	100	65	28	22
	220	120	58	53
	300	210	120	110
	430	280	230	140
	640	500	310	320
	1600	1100	500	1300
PPCC p-value	0.05	0.06	0.50	0.005

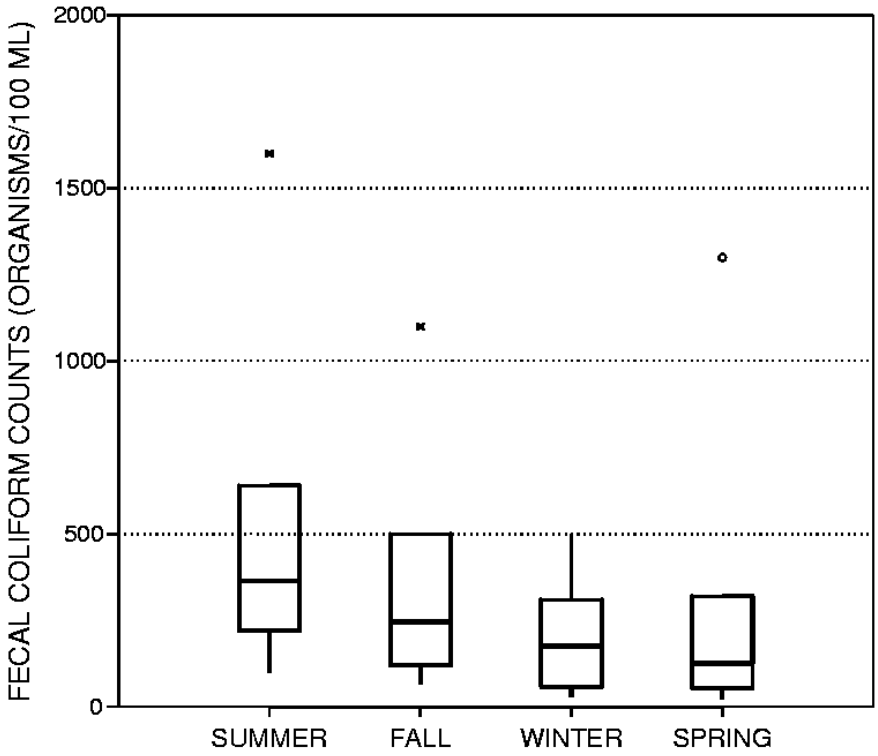


Figure 7.1 Boxplots of Fecal Coliform Data from the Illinois River

Should a parametric or nonparametric test be performed on these data? If even one of the four groups exhibits non-normality, the assumptions of parametric analysis of variance are violated. The consequences of this violation is an inability to detect differences which are truly present -- a lack of power. The PPCC test for normality rejects normality at α =0.05 for two of the seasons, summer and spring (table 7.1). Outliers and skewness for the fall samples also argue for non-normality. Based solely on the skewness and outliers evident in the boxplot, a nonparametric test should be used on these data.

Computation of the Kruskal-Wallis test is shown in table 7.2. This is compared to a table of the chi-square distribution available in many statistics texts, such as Iman and Conover (1983). We conclude that there is not enough evidence in these data to reject the assumption that fecal coliform counts are distributed similarly in all four seasons.

Table 7.2 Kruskal-Wallis test for the fecal coliform data.

	Summer	Fall	Winter	Spring	
Ranks R_{ij}	6	5	2	1	
	12	8.5	4	3	
	15	11	8.5	7	
	18	14	13	10	
	21	19.5	16	17	
	24	22	19.5	23	
$\overline{R}_j$	16	13.3	10.5	10.2	$\overline{R}_j$ = 12.5

K=2.69 $\chi^2_{0.95,(3)}$ = 7.815 p=0.44 so, do not reject equality of distributions.

7.1.1.4 The rank transform approximation

The rank transform approximation to the Kruskal-Wallis test is computed by performing a one-factor analysis of variance on the ranks R_{ij}. This approximation compares the mean rank within each group to the overall mean rank, using an F-distribution for the approximation of the distribution of K. The F and chi-square approximations will result in very similar p-values. The rank transform method should properly be called an "analysis of variance on the ranks".

For the example 1 data, the rank transform approximation results in a p-value of 0.47, essentially identical to that for the large sample approximation. Detailed computations are shown following the discussion of ANOVA in the next section.

7.1.2 Analysis of Variance (One Factor)

Analysis of variance is the parametric equivalent to the Kruskal-Wallis test. It compares the mean values of each group with the overall mean for the entire data set. If the group means are dissimilar, some of them will differ from the overall mean, as in figure 7.2. If the group means are similar, they will also be similar to the overall mean, as in figure 7.3.

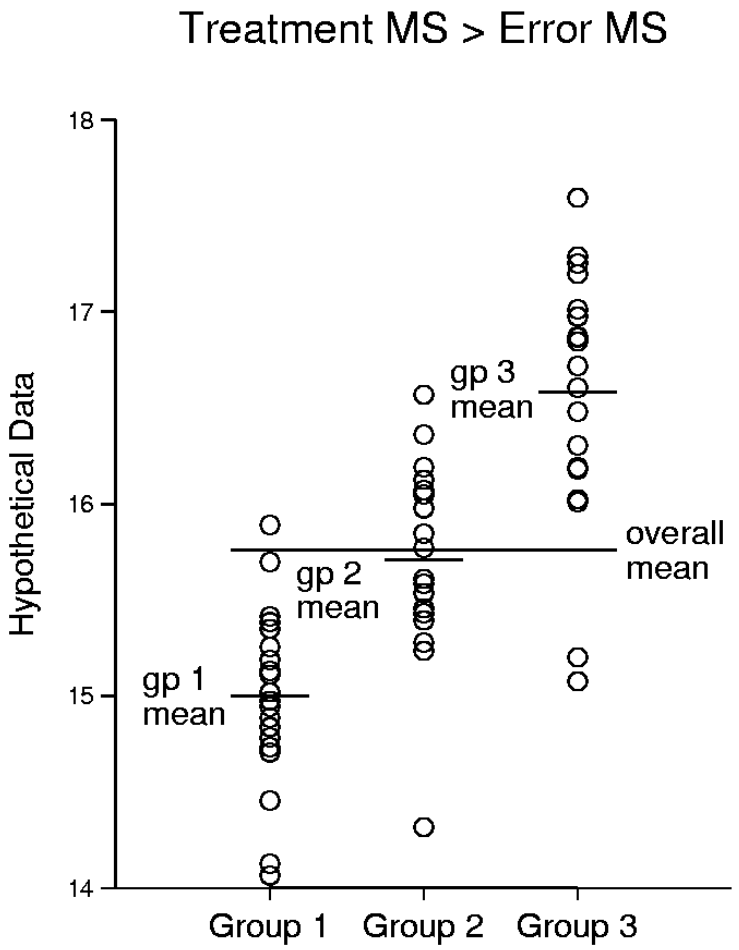


Figure 7.2 Hypothetical data for three groups.
Treatment mean square > Error mean square.

Why should a test of differences between means be named an analysis of variance? In order to determine if the differences between group means (the signal) can be seen above the variation within groups (the noise), the total noise in the data as measured by the total sum of squares is split into two parts:

Total sum of squares	=	**Treatment sum of squares**	+	**Error sum of squares**
(overall variation)	=	(group means – overall mean)	+	(variation within groups)
$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2$	=	$\sum_{j=1}^{k} n_j(\bar{y}_j-\bar{y})^2$	+	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$

If the total sum of squares is divided by N−1, where N is the total number of observations, it equals the variance of the y_{ij}'s. Thus ANOVA partitions the variance of the data into two parts, one measuring the signal and the other the noise. These parts are then compared to determine if the means are significantly different.

7.1.2.1 Null and alternate hypotheses

The null and alternate hypotheses for the analysis of variance are:

H_0: the k group means are identical $\mu_1 = \mu_2 = ... = \mu_k$.

H_1: at least one mean is different.

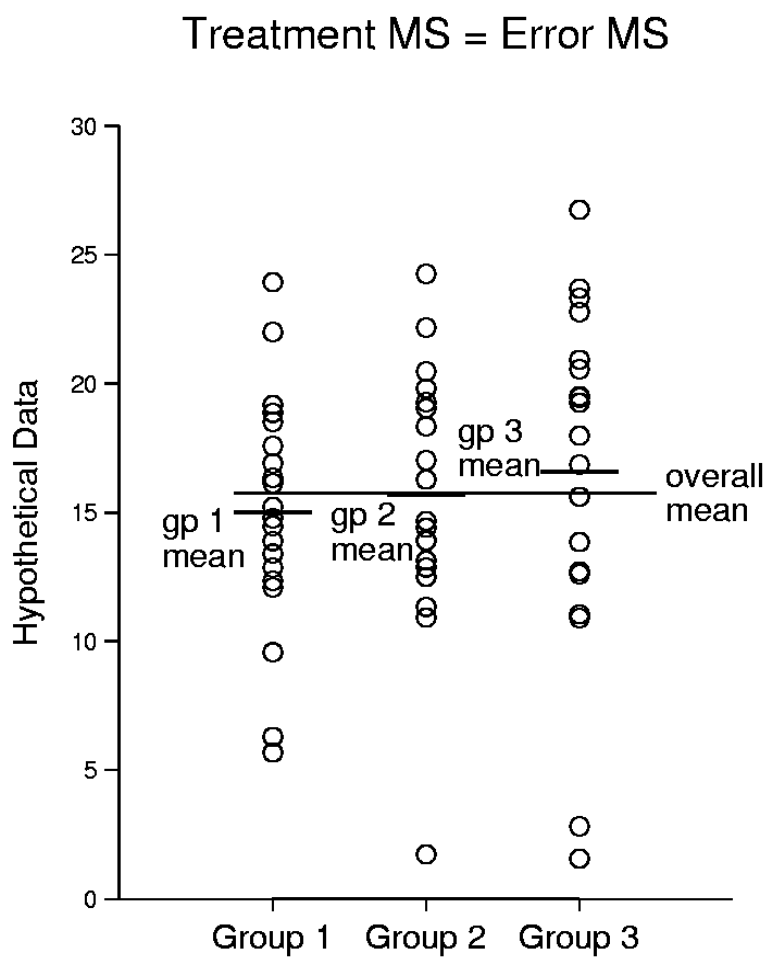


Figure 7.3 Hypothetical data for three groups.
Treatment mean square ≅ Error mean square.

7.1.2.2 Assumptions of the test

If ANOVA is performed on two groups, the F statistic which results will equal the square of the two-sample t-test statistic $F=t^2$, and will have the same p-value. It is not surprising, then, that the same assumptions apply to both tests:

1. All samples are random samples from their respective populations.
2. All samples are independent of one another.
3. Departures from the group mean ($y_{ij} - \bar{y}_j$) are normally distributed for all j groups.
4. All groups have equal population variance σ^2 estimated for each group by s_j^2

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{n_j - 1}$$

Violation of either the normality or constant variance assumption results in a loss of ability to see differences between means (a loss of power). The analysis of variance suffers from the same five problems as did the t-test: 1) lack of power when applied to non-normal data, 2) dependence on an additive model, 3) lack of applicability for censored data, 4) assumption that the mean is a good measure of central tendency for skewed data, and 5) difficulty in assessing whether the normality and equality of variance assumptions are valid for small sample sizes. See Chapter 5 for a detailed discussion of these problems.

Difficulties arise when using prior tests of normality to "prove" non-normality before allowing use of the nonparametric Kruskal-Wallis test. Small samples sizes may inhibit detecting non-normality, as mentioned above. Second, transformations must be done on more than two groups of data. It is usually quite difficult to find a single transformation which when applied to all groups will result in each becoming normal with constant variance. Even the best transformation based on sample data may not alleviate the power loss inherent when the assumptions of ANOVA are violated. Finally, if all groups are actually from a normal distribution, one or more may be "proven" non-normal simply by chance (there is an α% chance for each group). Thus the results of testing for normality can be quite inconclusive prior to performing ANOVA. The value of nonparametric approaches here is that they are relatively powerful for a wide range of situations.

7.1.2.3 Computation

Each observation y_{ij} can be written as a sum of the overall true mean μ, plus the difference α_j between μ and the true mean of the jth group μ_j, plus the difference ε_{ij} between the individual observation y_{ij} and the jth group mean μ_j:

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij},$$

where: y_{ij} is the ith individual observation in group j, j=1,...k;
μ is the overall mean (over all groups);
α_j is the "group effect", or $(\mu_j - \mu)$, and
ε_{ij} are the residuals or "error" within groups.

If H_0 is true, all j groups have the same mean equal to the overall mean μ, and thus $\alpha_j = 0$ for all j. If group means differ, $\alpha_j \neq 0$ for some j. In order to detect a difference between means, the variation within a group around its mean due to the ε_{ij}'s must be sufficiently small in comparison to the difference between group means so that the group means may be seen as different (see figure 7.2). The variation within a group is estimated by the within-group or error mean square (MSE), computed from the data. The variation between group means is estimated by the treatment mean square (MST). Their computation is shown below.

Sum of Squares

The error or within-group sum of squares

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

estimates the total within-group noise using departures from the sample group mean $\bar{y}_j$. Error in this context refers not to a mistake, but to the inherent variability within a group. The treatment (between-group) sum of squares

$$SST = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$

estimates the treatment effect using differences between group means and the overall mean of the sample, weighted by sample size.

Degrees of freedom

Each of the sums of squares has an associated degrees of freedom, the number of independent pieces of information used to calculate the statistic. For the treatment sum of squares this equals k−1, as when k−1 of the group means are known, the kth group mean can be calculated. The total sum of squares has N−1 degrees of freedom, the denominator of the formula for the variance of y_{ij}. The error sum of squares has degrees of freedom equal to the difference between the above two, or N−k.

Mean Squares and the F-test

Dividing the sums of squares by their degrees of freedom produces the total variance, and the mean squares for treatment (MST) and error (MSE). These mean squares are also measures of the variance of the data.

Mean Square		Formula	Estimates:
Variance of y_{ij}	=	Total SS / N−1	Total variance of the data
MST	=	SST / k−1	Variance within groups + variance between groups.
MSE	=	SSE / N−k	Variance within groups.

If H_0 is true, there is no variance between group means (no difference between means), and the MST will on average equal the MSE (figure 7.3). As $\alpha_j = 0$, all variation is simply around the overall mean μ, and the MST and MSE both estimate the total variance. However when H_1 is true, the MST is larger on average than the MSE (figure 7.2), as most of the noise is that between groups. Therefore a test is constructed to compare these two estimates of variance, MST and MSE. The F-ratio

$$F = MST / MSE$$

is computed and compared to quantiles of an F distribution. If MST is sufficiently larger than MSE, F is large and H_0 is rejected. When H_0 is true and there is no evidence for differences in group means, F is expected to equal 1 ($\mu_F = 1$ when H_0 is true). In other words, an F = 1 has a p-value near 0.50, varying with the degrees of freedom. If F were below 1, which could happen due to random variation in the data, generally $p > 0.50$ and no evidence exists for differences between group means.

The computations and results of an ANOVA are usually organized into an ANOVA table. For a one-way ANOVA, the table looks like:

Source	df	SS	MS	F	p-value
Treatment	(k−1)	SST	MST	MST/MSE	p
Error	(N−k)	SSE	MSE		
Total	N−1	Total SS			

Example 1, cont.

For the fecal coliform data from the Illinois River, the ANOVA table is given below. The F statistic is quite small, indeed below 1. At α=0.05 or any reasonable α-level, the mean counts would therefore not be considered different between seasons.

Source	df	SS	MS	F	p-value
Season	3	361397	120466	0.67	0.58
Error	20	3593088	179654		
Total	23	3954485			

However, this ANOVA has been conducted on non-normal data. Without knowing the results of the Kruskal-Wallis test, concern should be expressed that the result of "no difference" may be an artifact of the lack of power of the ANOVA, and not of a true equivalence of means. Some

statisticians have recommended performing both tests. This may be unnecessary if the data exhibit sufficient non-normality to suspect an inability of ANOVA to reject. Also assumed by performing ANOVA is that group means are an appropriate data summary. For the obviously skewed distributions found for all but the winter season, means will make little sense as estimates of the values which might be expected to occur. Means would be useful when estimating the mass of bacteria transported per season, but not in the hypothesis testing realm.

One factor analysis of variance

Situation Several groups of data are to be compared, to determine if their means are significantly different. Each group is assumed to have a normal distribution around its mean. All groups have the same variance.

Computation The treatment mean square and error mean square are computed as their sum of squares divided by their degrees of freedom (df). When the treatment mean square is larger than the error mean square as measured by an F-test, the group means are significantly different.

$$\text{MST} = \frac{\sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2}{k-1}$$ where $k-1$ = treatment degrees of freedom

$$\text{MSE} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{N-k}$$ where $N-k$ = error degrees of freedom

Tied data No alterations necessary.

Test Statistic The test statistic F:
F = MST / MSE

Decision Rule To reject H_0: the mean of every group is identical, versus
H_1: at least one mean differs .
Reject H_0 if $F \geq F_{1-\alpha,\, k-1,\, N-k}$ the $1-\alpha$ quantile of an F distribution with $k-1$ and $N-k$ degrees of freedom; otherwise do not reject H_0.

7.2 Tests for the Effects of More Than One Factor

It is quite common that more than one factor is suspected to be influencing the magnitudes of observations. In these situations it is desirable to measure the influence of all factors simultaneously. Sequential one-factor tests are an inadequate alternative to a single multi-factor

test. Even when only one factor is actually influencing the data and a one-way ANOVA for that factor soundly rejects H_0, a second one-way test for a related factor may erroneously reject H_0 simply due to the association between the two factors. The test for the second factor should remove the effect of the first before establishing that the second has any influence. By evaluating all factors simultaneously, the influence of one can be measured while compensating for the others. This is the objective of a multi-factor analysis of variance, and of the nonparametric analogue.

7.2.1 Nonparametric Multi-Factor Tests

For two-factor and more complex ANOVA's where the data within one or more treatment groups are not normally distributed and may not have equal variances, there are two possible approaches for analysis. The first is a class of tests which include the Kruskal-Wallis and Friedman tests as simpler cases. These tests, described by Groggel and Skillings (1986), do not allow for interactions between factors. The tests reformat multiple factors into two factors, one the factor being tested, and the other the collection of all other treatment groups for all remaining factors. The data are then ranked within treatment groups for analysis, much as in a Friedman test. The reader is referred to their paper for more detail.

The second procedure is a rank transformation test (Conover and Iman, 1981). All data are ranked from 1 to N, and an ANOVA computed on the ranks. This procedure is far more robust to departures from the assumptions of normality and constant variance than is an ANOVA on the original data. The rank transformation produces values which are much closer to meeting the two critical assumptions than are the original values themselves. The tests determine whether the mean rank differs between treatment groups, rather than the mean. The mean rank is interpreted as an estimate of the median. Multiple comparison procedures on the ranks can then differentiate which groups differ from others.

Examples of the computation and performance of these rank transformation tests will be delayed until after discussion of parametric factorial ANOVA.

7.2.2 Multi-Factor Analysis of Variance -- Factorial ANOVA

The effects of two or more factors may be simultaneously evaluated using a factorial ANOVA design. A factorial ANOVA occurs when none of the factors is a subset of the others. If subsetted factors do occur, the design includes "nested" factors and the equations for computing the F test statistics will differ from those here (nested ANOVA is briefly introduced in a later section). A two-factor ANOVA will be fully described -- more than two factors can be incorporated, but are beyond the scope of this book. See Neter, Wasserman and Kutner (1985) for more detail on higher-way and nested analysis of variance.

For a two-factor ANOVA, the influences of two explanatory variables are simultaneously tested. The first page of this chapter presented a two-factor ANOVA, the determination of chemical concentrations among basins at low flow. The objective was to determine whether concentrations differed as a function of mining history (whether or not each basin was mined, and if so whether it was reclaimed) and of rock type.

7.2.2.1 Null and alternate hypotheses

Call the two factors A and B. There are $i=1,...a \geq 2$ categories of factor A, and $j=1,...b \geq 2$ categories of factor B. Treatment groups are defined as all the possible combinations of factors A and B, so there are a•b treatment groups. Within each treatment group there are n_{ij} observations. The test determines whether mean concentrations are identical among all the a•b treatment groups, or whether at least one differs.

H_0 : all a•b treatment group means μ_{ij} are equal $\mu_{11} = \mu_{12} = ... = \mu_{ab}$

H_1 : at least one μ_{ij} differs from the rest.

The magnitude of any observation y_{ijk} can be affected by several possible influences:

$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$, where

α_i = influence of the ith category of A

β_j = influence of the jth category of B

$\alpha\beta_{ij}$ = interaction effect between A and B beyond those of α_i and β_j separately for the ijth treatment group, and

ε_{ijk} = residual error, the difference between the kth observation ($k=1,...n_{ij}$) and the treatment group mean μ_{ij}.

The null hypothesis states that treatment group means μ_{ij} all equal the overall mean μ. Therefore α_i β_j and $\alpha\beta_{ij}$ all equal 0 -- there are no effects due to any of the factors or to their interaction. If any one of α_i, β_j, or $\alpha\beta_{ij}$ are nonzero, the null hypothesis is rejected, and at least one treatment group evidences a difference in its mean.

7.2.2.2 Interaction between factors

If $\alpha\beta_{ij} = 0$ in the equation above, there is no interaction present. Without interaction, the effect of factor B is identical for all groups of factor A, and the effect of factor A is identical for all groups of factor B. Suppose there are 3 groups of factor A (a1, a2, and a3) and 2 groups of factor B (b1 and b2), resulting in six treatment groups overall. Lack of interaction can be visualized by plotting the means for all treatment groups as in figure 7.4. The parallelism of the lines shows that no interaction is present. The effect of A going from a1 to a2 to a3 is identical regardless of which B group is involved. The increase going from b1 to b2 for factor B is identical for every group of factor A.

When interaction is present ($\alpha\beta_{ij} \neq 0$) the treatment group means are not determined solely by the additive effects of factors A and B alone. Some of the groups will have mean values larger

or smaller than those expected just from the results of the individual factors. The effect of factor A can no longer be discussed without reference to which group of factor B is of interest, and the effect of factor B can likewise not be stated apart from a knowledge of the group of factor A. In a plot of the treatment group means, the lines are no longer parallel (figure 7.5). The pattern of differences going from a1 to a2 to a3 depends on which group of factor B is of interest, and likewise for the differences between b1 and b2 -- the pattern differs for the three A groups.

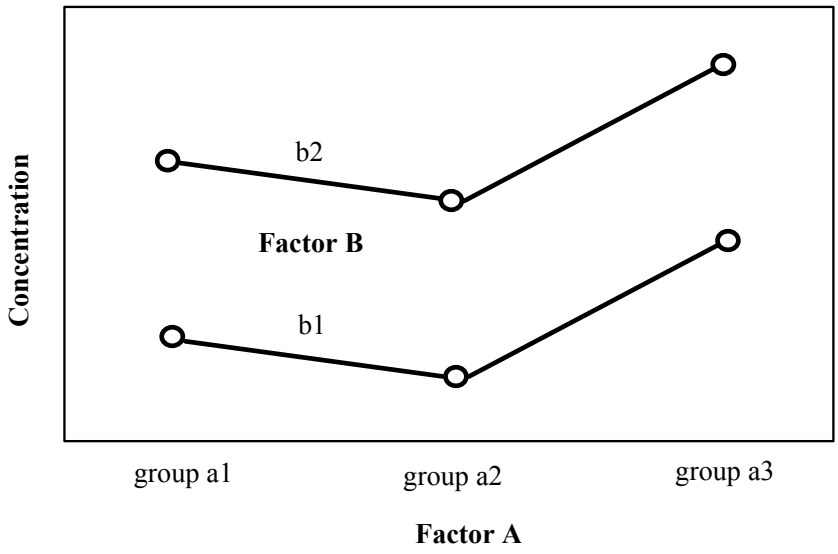


Figure 7.4 Six treatment group means with no interaction present

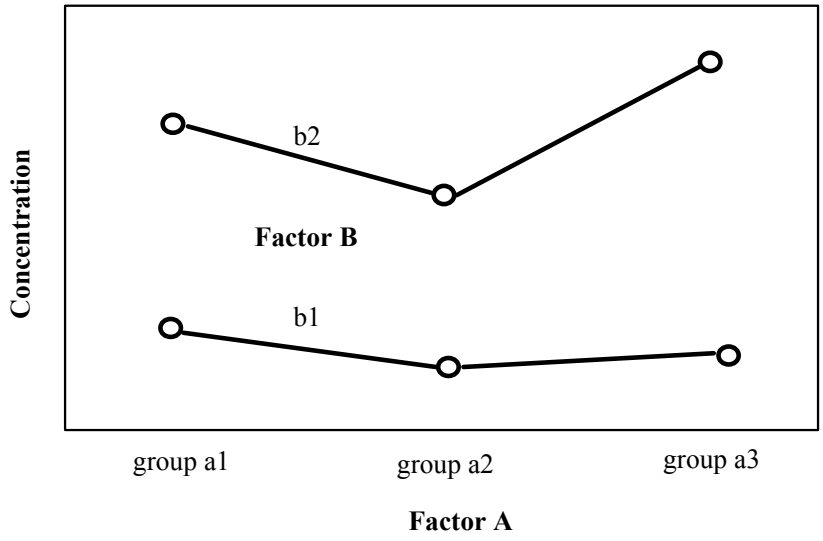


Figure 7.5 Six treatment group means with interaction present

Interaction can result from a synergistic or antagonistic effect. As an example, fish may not die instream due only to higher water temperatures, or to slightly higher copper concentrations, but combine the two and the result could be deadly. This type of interrelation between factors results in a significant interaction effect. For k factors there are (k–1) possible interaction terms between the factors. Unless it is known ahead of time that interactions are not possible, interaction terms should always be included and tested for in multi-factor ANOVA models.

7.2.2.3 Assumptions for factorial ANOVA

Assumptions are the same as for a one-way ANOVA. Departures from each treatment group mean μ_{ij} (every combination of factors A and B) are assumed normally distributed with identical variance. This is a consequence of the ε_{ij}, which are normally distributed and of variance σ^2, being randomly distributed among the treatment groups. The normality and constant variance assumptions can be checked by inspecting boxplots of the data for each treatment group.

7.2.2.4 Computation

The influences of factors A, B, and their interaction are evaluated separately by again partitioning the total sums of squares into component parts due to each factor. After dividing by their respective degrees of freedom, the mean squares for factors A, B, and interaction are produced. As with a one-way ANOVA, these are compared to the error mean square (MSE) using F-tests to determine their significance.

<u>Sum of Squares</u>

The equations for the sums of squares for factor A (SSA), factor B (SSB), interaction (SSI), and error, assuming constant sample size n per treatment group, are:

		due to
SSA =	$\sum^{a} \frac{(\sum^{b}\sum^{n} y)^2}{bn} - \frac{(\sum^{a}\sum^{b}\sum^{n} y)^2}{abn}$	$\mu_i - \mu$
SSB =	$\sum^{b} \frac{(\sum^{a}\sum^{n} y)^2}{an} - \frac{(\sum^{a}\sum^{b}\sum^{n} y)^2}{abn}$	$\mu_j - \mu$
SSI =	Total SS – SSA – SSB – SSE	$\mu_{ij} - (\mu_i + \mu_j) + \mu$
SSE =	$\sum^{a}\sum^{b}\sum^{n} (y)^2 - \sum^{a}\sum^{b} \frac{(\sum^{n} y)^2}{n}$	$y_{ijk} - \mu_{ij}$
Total SS =	$\sum^{a}\sum^{b}\sum^{n} (y)^2 - \frac{(\sum^{a}\sum^{b}\sum^{n} y)^2}{abn}$	$y_{ijk} - \mu$

Mean Squares and the F-test

Dividing the sums of squares by their degrees of freedom produces the mean squares for factors A, B, interaction, and error as in the ANOVA table below. If H_0 is true and α_i, β_j, and $\alpha\beta_{ij}$ all equal 0, all variation is simply around the overall mean μ. The MSA, MSB, and MSI will then all be measures of the error variance, as is the MSE, and all three F-tests will have ratios not far from 1. However when H_1 is true, at least one of the mean squares in the numerators should be larger than the MSE, and the resulting F-ratio will be larger than the appropriate quantile of the F distribution. When F is large, H_0 can be rejected, and that influence be considered to significantly affect the magnitudes of the data at a level of risk equal to α.

The two-factor ANOVA table is as follows when there is an equal number of observations for each treatment (all $n_{ij} = n$).

Source	df	SS	MS	F	p-value
Factor A	(a−1)	SSA	SSA/(a-1)	MSA/MSE	
Factor B	(b−1)	SSB	SSB/(b-1)	MSB/MSE	
Interaction	(a−1) (b−1)	SSI	SSI/(a-1)(b-1)	MSI/MSE	
Error	ab(n−1)	SSE	SSE/[ab(n-1)]		
Total	abn−1	Total SS			

Multi-factor analysis of variance

Situation Two or more influences are to be simultaneously tested, to determine if either cause significant differences between treatment group means. Each group is assumed to have a normal distribution around its mean. All groups have the same variance.

Computation Compute the sums of squares and mean squares as above.

Tied data No alterations necessary.

Test Statistic

To test factor A:	To test factor B:	To test for interaction:
F_A = MSA / MSE	F_B = MSB / MSE	F_I = MSI / MSE

with degrees of freedom for the numerator of:

dfn = (a−1)	dfn = (b−1)	dfn = (a−1)(b-1)

Decision Rule To reject H_0: the mean of every group is identical (no treatment effects for either factor or interaction), versus

H_1: at least one mean differs.

Reject H_0 if $F \geq F_{1-\alpha,\, dfn,\, ab(n-1)}$ the $1-\alpha$ quantile of an F distribution with dfn and ab(n−1) degrees of freedom; otherwise do not reject H_0.

Example 2

Iron concentrations were measured at low flow in numerous small streams in the coal-producing areas of eastern Ohio (Helsel, 1983). Each stream drains either an unmined area, a reclaimed coal mine, or an abandoned coal mine. Each site is also underlain by either a sandstone or limestone formation. Are iron concentrations influenced by upstream mining history, by the underlying rock type, or by both?

There are several scenarios which would cause H_0 to be rejected. Factor A (say mining history) could be significant ($\alpha_i \neq 0$), but factor B insignificant. Or factor B (rock type) could be significant ($\beta_j \neq 0$), but not A. Both factors could be significant ($\alpha_i, \beta_j \neq 0$). Both factors could be significant, plus an additional interaction effect because one or more treatment groups (say unreclaimed sandstone basins) exhibited much different iron concentrations than those expected from either influence alone ($\alpha_i, \beta_j, \alpha\beta_{ij} \neq 0$). Finally, both factor A and B could be not significant ($\alpha_i, \beta_j = 0$) but concentrations be elevated for one specific treatment group ($\alpha\beta_{ij} \neq 0$). This would be interpreted as no overall mining or rock type effect, but one combination of mining history and rock type would have differing mean concentrations.

Boxplots for a subset of the iron concentration data from Helsel (1983) are presented in figure 7.6. Note the skewness, as well as the differences in variance as depicted by differing box heights. A random subset was taken in order to produce equal sample sizes per treatment group, yet preserving the essential data characteristics. The subset data are listed in Appendix C5. In the section 7.2.2.5, analysis of unequal sample sizes per treatment group will be presented and the entire iron data set analyzed.

There are six treatment groups, combining the three possible mining histories (unmined, abandoned mine, and reclaimed mine) and the two possible rock types (sandstone and limestone). An analysis of variance conducted on this subset which has n=13 observations per treatment group produced the following ANOVA table. Tested was the effect of mining history alone, rock type alone, and their interaction (Mine*Rock). A*B is a common abbreviation for the interaction between A and B.

ANOVA table for the subset of iron data

Source	df	SS	MS	F	p-value
Rock	1	15411	15411	2.38	0.127
Mine	2	32282	16141	2.49	0.090
Rock*Mine	2	25869	12934	2.00	0.143
Error	72	466238	6476		
Total	77	539801			

None of the three possible influences is significant at the $\alpha = 0.05$ level, as their p-values are all larger than 0.05. However, the gross violation of the test's assumptions of normality and equal

variance shown in the boxplots must be considered. Perhaps the failure to reject H_0 is due not to a lack of an influence evidenced in the data, but of the parametric test's lack of power to detect these influences because of the violation of test assumptions. To determine whether this is so, the equivalent rank transformation test is performed.

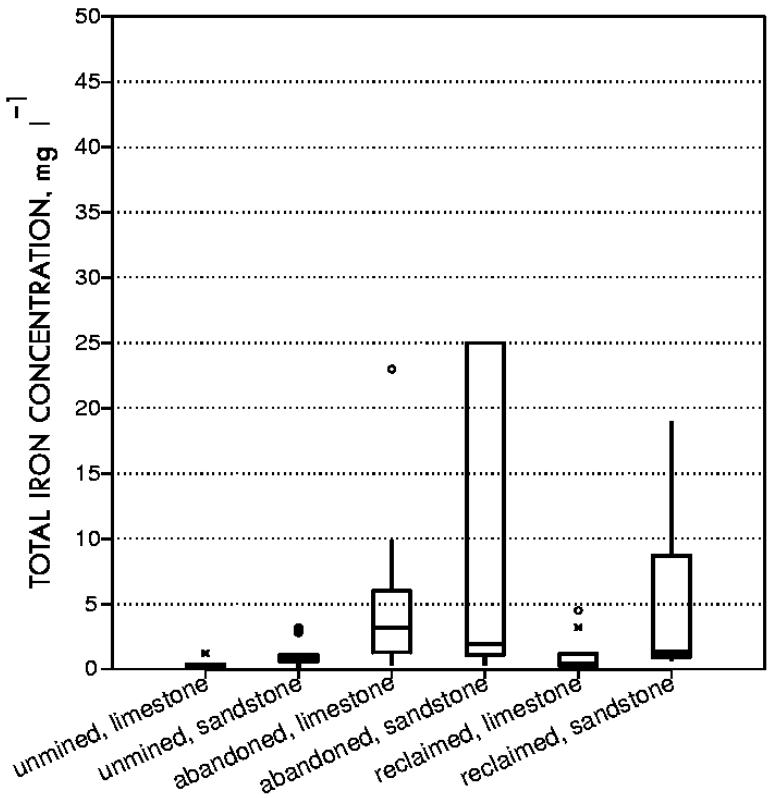


Figure 7.6 A subset of the iron concentrations at low flow from Helsel (1983)

To compute the rank transformation test, the data are ranked from smallest to largest, 1 to n=78. An analysis of variance is then performed on the ranks of the data. The ANOVA table is below, while a boxplot of data ranks is shown in figure 7.7.

ANOVA table for the ranks of the subset of iron data

Source	df	SS	MS	F	p-value
Rock	1	4121.7	4121.7	13.38	0.000
Mine	2	10933.9	5467.0	17.74	0.000
Rock*Mine	2	2286.2	1143.1	3.71	0.029
Error	72	22187.2	308.2		
Total	77	39529.0			

Results for the rank transformation tests are startlingly different than those for the parametric ANOVA. All three influences, mining history, rock type, and their interaction, are significant at $\alpha = 0.05$. Gross violations of the assumptions of ANOVA by these data have clearly inhibited the parametric test from detecting the influences of these factors. The rejection of H_0 for the rank test indicates that the median iron concentrations differ between treatment groups. Mean concentrations will be distorted by the skewness and outliers present in most of the treatment groups.

Analysis of variance on data ranks is an "asymptotically distribution-free" technique. That is, for sufficiently large sample sizes it tests hypotheses which do not require the assumption of data normality. For the cases where equivalent, truly nonparametric techniques exist such as the Kruskal-Wallis and Friedman tests, the rank transformation procedures have been shown to be large-sample approximations to the test statistics for those techniques. Where no equivalent nonparametric methods have yet been developed such as for the two-way design, rank transformation results in tests which are more robust to non-normality, and resistant to outliers and non-constant variance, than is ANOVA without the transformation.

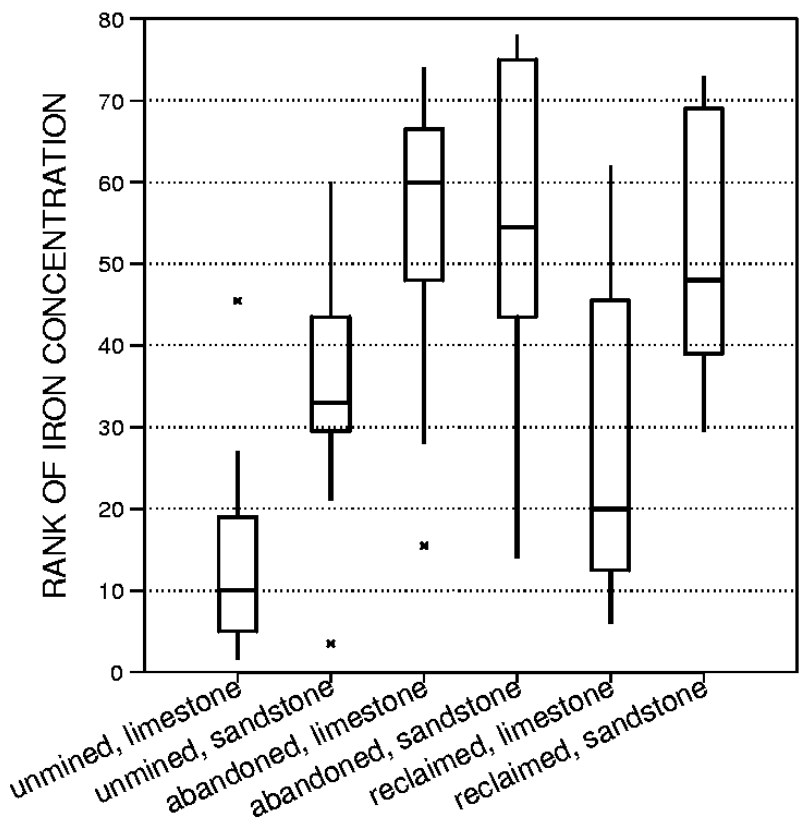


Figure 7.7 Boxplots of the ranks of the iron data shown in Figure 7.6

A third option for analysis of the two-way design is ANOVA on data transformed by a power transformation. The purpose of the power transformation is to produce a more nearly-normal

and constant variance data set. As water resources data are usually positively skewed, the log transformation is often employed. Using logarithms for ANOVA implies that the influences of each factor are multiplicative in the original units, as the influences of the logarithms are additive. The primary difficulty in using a power transformation is in producing a normally distributed error structure for every treatment group. Groups which are skewed may be greatly aided by a transformation, but be side-by-side with a group which was symmetric in the original units, and is now asymmetric after transformation! Boxplots for each treatment group should be inspected prior to performing the ANOVA to determine if each group is at least symmetric. When only some of the treatment groups exhibit symmetry, much less normality, concerns over the power of the procedure remain. F tests which appear to be not significant are always suspect.

In figure 7.8, boxplots of the base 10 logarithms of the low-flow iron concentrations are presented. Most of the treatment groups still remain distinctly right-skewed even after the transformation, while the unmined limestone group appears less symmetric following transformation! There is nothing magic in the log transformation -- any other transformation going down the ladder of powers might also remedy positive skewness. It may also alter a symmetric group into one that is left-skewed. The search for a transformation which results in all groups being symmetric is often fruitless. In

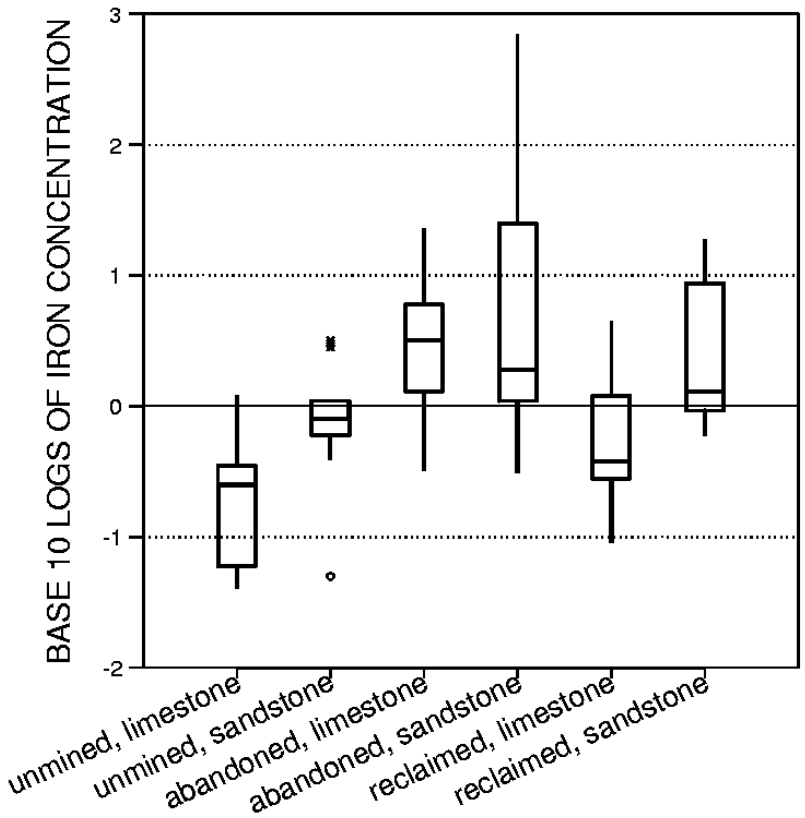


Figure 7.8 Boxplots of the base 10 logarithms of the iron data shown in Figure 7.6

addition, the "best" power transformation will likely change going from one data set to another, one location to another, and one time period to another. In comparison, the rank transformation has simplicity, comparability among locations and time periods, and general validity as being asymptotically distribution-free. When the assumptions of normality and constant variance are questionable, the rank transformation is the most generally appropriate alternative.

7.2.2.5 Unequal sample sizes

Equations presented in the previous section are appropriate only when the number of data per treatment group is identical for each group. This is also called a "balanced" design. Computations for unequal sample sizes ("unbalanced" designs) are more complex. Smaller statistics software packages often encode tests valid only for balanced designs, though that is not always obvious from their output. Yet water resources data rarely involve situations when all sample sizes are equal. Sample bottles are broken, floods disrupt the schedule, etc. When data are unbalanced, the sums of squares for the above equations no longer test

$$H_0: \mu_1 = \mu_2 = ... = \mu_k$$

but test instead an hypothesis involving weighted group means, where the weights are a function of treatment group sample sizes. This is of little use to the practitioner. Some software will output the (useless and incorrect) results valid only for equal sample sizes even when unbalanced data are provided as input, with no warnings of their invalidity. Be sure that when unequal sample sizes occur, tests which can incorporate them are performed.

To perform ANOVA on unbalanced data, a regression approach is necessary. This is done on larger statistical packages such as Minitab or SAS. SAS's "type I" sums of squares (called "sequential sums of squares" by Minitab) are valid only for balanced cases, but SAS's "type III" sums of squares (Minitab's "adjusted sums of squares") are valid for unbalanced cases as well. Unbalanced ANOVAs are computed in the same fashion as nested F-tests for comparing regression models in analysis of covariance, discussed in Chapter 11. Because the equations for the sums of squares are "adjusted" for unequal sample sizes, they do not sum to the total sum of squares as for balanced ANOVA. See Neter, Wasserman and Kutner (1985) for more detail on the use of regression models for performing unbalanced ANOVA.

<u>Example 2, continued</u>

The complete 241 observations (Appendix C6) from Helsel (1983) are analyzed with an unbalanced ANOVA. Boxplots for the six treatment groups are shown in figure 7.9. They are quite similar to those in figure 7.6, showing that the subsets adequately represented all the data. An ANOVA table for the complete iron data set is as follows. Note that the sums of squares do not add together to equal the total sum of squares for this unbalanced ANOVA. Results for these data would be incorrect if performed by software capable only of balanced ANOVA. Conclusions reached (do not reject for all tests) agree with those previously given for ANOVA on the data subset.

ANOVA table for the complete (unbalanced) iron data

Source	df	SS	MS	F	p-value
Rock	1	71409	71409	0.51	0.476
Mine	2	262321	131160	0.93	0.394
Rock*Mine	2	178520	89260	0.64	0.530
Error	235	32978056	140332		
Total	240	34062640			

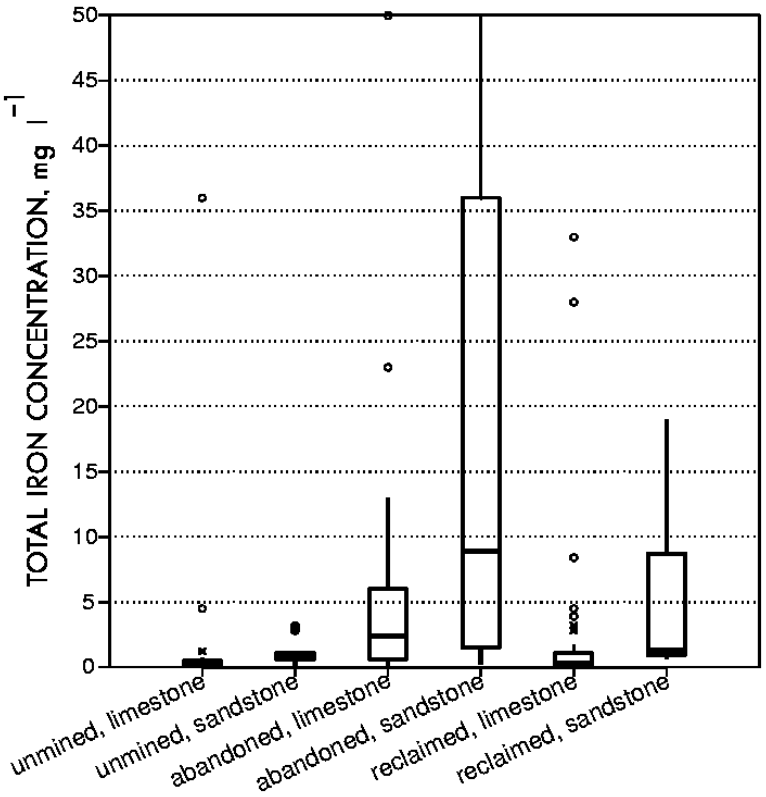


Figure 7.9 Iron concentrations at low flow from Helsel (1983)

7.2.2.6 Fixed and random factors

An additional requirement for the F tests previously given is that both factors are fixed. With a fixed factor, the inferences to be made from the results extend only to the treatment groups under study. For example, the influences of unmined, abandoned, and reclaimed mining histories were previously compared. Differences in resulting chemical concentrations between these three specific mining histories are of interest, and hence this is a fixed factor. A random factor would result from a random selection of several groups out of a larger possible set to represent the overall factor. Inferences from the test results would be extended beyond the specific groups being tested to the generic factor itself. Thus there is little or no interest in attributing test results to a specific individual group, but only in ascertaining a generic effect due to that factor.

As an example, suppose soil concentrations of a trace metal are to be compared between three particle size fractions all across the state, to determine which of the three fractions is most appropriate as a reconnaissance medium. Particle size is a fixed effect -- there is interest in those specific sizes. However, there is only enough funding to sample sparsely if done all across the state, so instead a random factor is incorporated to determine whether spatial differences occur. Several counties are selected at random, and intensive sampling occurs within those counties. No sampling is done outside of those counties. The investigator will determine not only which size fraction is best, but whether this is consistent among the counties (the random effect), which by inference is extended to the entire state. There is no specific interest in the counties selected, but only as they represent spatial variability.

If every factor were random, F tests would use the mean squares for interaction as denominators rather than the mean square for error. If a mix of random and fixed factors occurs (called a "mixed effects" design) as in the example above, there would be a mixture of mean squares used as denominators. In general the fixed factors in the design use the interaction mean squares as denominators, and the random factors the error mean square, the reverse of what one might intuitively expect! However, the structure of mixed effects F tests can get much more complicated, especially for more than two factors, and texts such as Neter, Wasserman and Kutner (1985) or Sokal and Rohlf (1981) should be consulted for the correct setup of F tests when random factors are present. Note that computer software uses the MSE in the denominator unless otherwise specified, and thus assumes that all factors are fixed. Therefore F tests automatically produced will not be correct when random factors are present, and the correct F ratio must be specifically requested and computed.

7.3 Blocking -- The Extension of Matched-Pair Tests

In Chapter 6, tests for differences between matched-pairs of observations were discussed. Each pair of observations had one value in each of two groups, such as "before" versus "after". The advantage of this type of design is that it "blocks out" the differences from one matched-pair to another that is contributing unwanted noise. Such noise may mask the differences between the two groups (the treatment effect being tested) unless matched-pairs are used.

Similar matching schemes can be extended to test more than two treatment groups. Background noise is eliminated by applying the treatment to blocks (rather than pairs) of similar or identical individuals. Only one observation is usually available for each combination of treatment and block. This is called a "randomized complete block design", and is a common design in the statistical literature.

The third example at the beginning of this chapter, detecting differences between three extraction methods used at numerous wells, is an example of this design. The treatment effect is

the extraction method, of which there are three types (three groups). The blocking effect is the well location; the well-to-well differences are to be "blocked out". One sample is analyzed for each extraction method at each well.

Four methods for analysis of a randomized complete block design will be presented. Each of them attempts to measure the same influences. To do this, each observation y_{ij} is broken down into the effects of four influences:

$$y_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij},$$

where y_{ij} is the individual observation in block i and group j;
μ is the overall mean or median (over all groups),
α_j is the "jth group effect", j=1,k
β_i is the "ith block effect", i=1,n
ε_{ij} is the residual or "error" between the individual observation and the combined group and block effects.

Median polish provides resistant estimates of the overall median, of group effects and block effects. It is an exploratory technique, not an hypothesis test procedure. Related graphical tools determine whether the two effects are additive or not, and whether the ε_{ij} are normal, as assumed by an ANOVA. If not, a transformation should be employed to achieve additivity and normality before an ANOVA is performed. The Friedman and median aligned ranks tests are nonparametric alternatives for testing whether the treatment effect is significant in the presence of blocking.

7.3.1 Median Polish

Median polish (Hoaglin et al., 1983) is an iterative process which provides a resistant estimate m of the overall median μ, as well as estimates a_j of the group effects α_j and b_i of the block effects β_i. Its usefulness lies in its resistance to effects of outliers. The polishing is begun by subtracting the medians of each row from the data table, leaving the residuals. The median of these row medians is then computed as the first estimate of the overall median, and subtracted from the row medians. The row medians are now the first estimates of the row effects. Then the median of each column is subtracted from the residual data table and set aside. The median of the column medians is subtracted from the column medians, and added to the overall median. The column medians now become the first estimates of the column effects. The entire process is repeated a second time, producing an estimated overall median m, row and column departures from the overall median (estimates a_j and b_i), and a data table of residuals eij estimating the ε_{ij}.

<u>Example 3</u>

Mercury concentrations were measured in periphyton at six stations along the South River, Virginia, above and below a large mercury contamination site (Walpole and Myers, 1985). Measurements were made on six different dates. Of interest is whether the six stations differ in mercury concentration. Is this a one-way ANOVA setup? No, because there may be

differences among the six dates -- the periphyton may not take up mercury as quickly during some seasons as others, etc. Differences caused by sampling on six different dates are unwanted noise which should be blocked out, hence date is a blocking effect. The data are presented in table 7.3, and boxplots by station in figure 7.10. There appears to be a strong increase in mercury concentration going downstream from station 1 to station 6, reflecting an input of mercury along the way.

Table 7.3 Mercury Concentrations in Periphyton (Walpole and Myers, 1985)

Station:	1	2	3	4	5	6
Date						
1	0.45	3.24	1.33	2.04	3.93	5.93
2	0.10	0.10	0.99	4.31	9.92	6.49
3	0.25	0.25	1.65	3.13	7.39	4.43
4	0.09	0.06	0.92	3.66	7.88	6.24
5	0.15	0.16	2.17	3.50	8.82	5.39
6	0.17	0.39	4.30	2.91	5.50	4.29

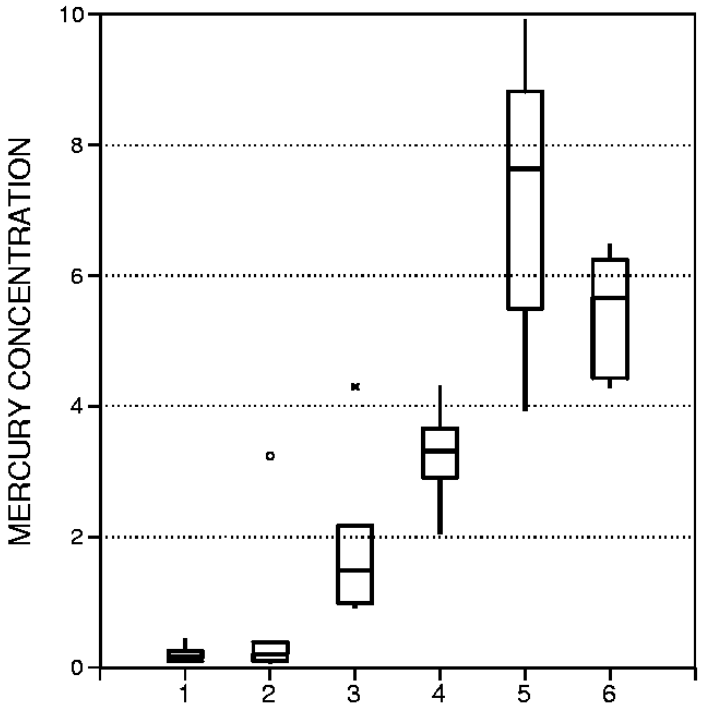


Figure 7.10 Periphyton Mercury Upstream (1) to Downstream (6) of Input to River

The first step in median polish is to compute the median of each row (date), and subtract it from that row's data. The residuals remain in the table.

Table 7.4 Table 7.3 data aligned by subtraction of row medians

Station:	1	2	3	4	5	6	row med
Date							(b_i)
1	-2.190	0.600	-1.310	-0.600	1.290	3.290	2.64
2	-2.550	-2.550	-1.660	1.660	7.270	3.840	2.65
3	-2.140	-2.140	-0.740	0.740	5.000	2.040	2.39
4	-2.200	-2.230	-1.370	1.370	5.590	3.950	2.29
5	-2.685	-2.675	-0.665	0.665	5.985	2.555	2.84
6	-3.430	-3.210	0.700	-0.690	1.900	0.690	3.60

Next the median of the row medians (2.64) is computed as the first estimate of the overall median m. This is subtracted from each of the row medians:

Station:	1	2	3	4	5	6	row med
Date							(b_i)
1	-2.19	0.60	-1.31	-0.60	1.29	3.29	0.00
2	-2.55	-2.55	-1.66	1.66	7.27	3.84	0.01
3	-2.14	-2.14	-0.74	0.74	5.00	2.04	-0.25
4	-2.20	-2.23	-1.37	1.37	5.59	3.95	-0.35
5	-2.69	-2.68	-0.67	0.67	5.99	2.56	0.20
6	-3.43	-3.21	0.70	-0.69	1.90	0.69	0.96
							m=**2.64**

The median of each column (station) is then computed and subtracted from that column's data. The residuals from the subtractions remain in the table.

Station:	1	2	3	4	5	6	row med
Date							(b_i)
1	0.19	2.99	-0.29	-1.31	-4.01	0.37	0.00
2	-0.17	-0.16	-0.64	0.95	1.97	0.92	0.01
3	0.24	0.25	0.28	0.03	-0.30	-0.88	-0.25
4	0.18	0.16	-0.35	0.66	0.29	1.03	-0.35
5	-0.31	-0.29	0.35	-0.04	0.69	-0.36	0.20
6	-1.05	-0.82	1.72	-1.40	-3.40	-2.23	0.96
a_j col med:	-2.38	-2.39	-1.02	0.71	5.30	2.92	m=**2.64**

Then the median of the column medians (-0.16) is subtracted from each of the column medians, and added to the overall median:

Station:	1	2	3	4	5	6	row med
Date							(b_i)
1	0.19	2.99	-0.29	-1.31	-4.01	0.37	0.00
2	-0.17	-0.16	-0.64	0.95	1.97	0.92	0.01
3	0.24	0.25	0.28	0.03	-0.30	-0.88	-0.25
4	0.18	0.16	-0.35	0.66	0.29	1.03	-0.35
5	-0.31	-0.29	0.35	-0.04	0.69	-0.36	0.20
6	-1.05	-0.82	1.72	-1.40	-3.40	-2.23	0.96
a_j col med:	-2.22	-2.23	-0.86	0.87	5.46	3.08	m=**2.48**

This table now exhibits the first "polish" of the data. Usually two complete polishes are performed in order to produce more stable estimates of the overall median and row and column effects. For the second polish, the above process is repeated on the table of residuals from the first polish. After a second complete polish, little change in the estimates is expected from further polishing. The table then looks like:

Station:	1	2	3	4	5	6	row med
Date							(b_i)
1	0.22	3.02	-0.19	-1.26	-3.77	0.31	0.03
2	-0.57	-0.56	-0.97	0.57	1.78	0.43	0.47
3	0.08	0.09	0.19	-0.11	-0.24	-1.12	-0.03
4	-0.08	-0.09	-0.54	0.42	0.24	0.69	-0.03
5	-0.17	-0.14	0.56	0.11	1.04	-0.31	0.12
6	0.15	0.38	2.99	-0.18	-1.98	-1.11	-0.18
a_j col med:	-2.18	-2.19	-0.89	0.89	5.29	3.20	m=**2.38**

The above table shows that

1) The station effects are large in comparison to the date effects (the a_j are much larger in absolute magnitude than the b_i).
2) There is a clear progression from smaller to larger values going downstream (a_j generally increases from stations 1 to 6), with the maximum at station 5.
3) A large residual occurs for station 5 at date 1 (smaller concentration than expected).

7.3.1.1 Plots related to median polish for checking assumptions

Median polish can be used to check the assumptions behind an analysis of variance. The first assumption is that the residuals ε_{ij} are normally distributed. Boxplots of the residuals e_{ij} in the table provide a look at the distribution of errors after the treatment and block effects have been removed. Figure 7.11 shows that for the periphyton mercury data the residuals are probably not normal due to the large proportion of outliers, but at least are relatively symmetric:

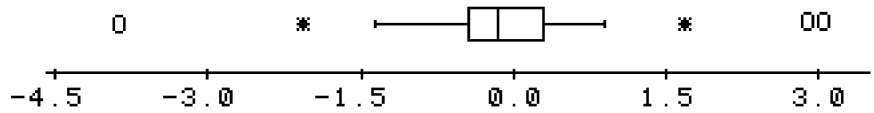


Figure 7.11 Residuals from the median smooth of periphyton mercury data

In addition, the additivity of the table can be checked. An ANOVA assumes that the treatment and block effects are additive. In other words, if being in group 1 adds -2.18 units of concentration to the overall mean or median, and if being at time 1 adds 0.03 units, these add together for treatment group 1 at time 1. If this is not the case, a transformation of the data prior to ANOVA must be performed to produce additivity. To check additivity, the "comparison value" c_{ij} (Hoaglin et al., 1983) is computed for each combination ij of block and treatment group, where

$$c_{ij} = a_i \bullet b_j / m .$$

A residuals plot of the tabled residuals e_{ij} versus c_{ij} will appear to have a random scatter around 0 if the data are additive. If not, the pattern of residuals will lead to an appropriate transformation to additivity -- for a nonzero slope s, the data should be raised to the (1−s) power in the ladder of powers. In figure 7.12, a residuals plot for the mercury median polish indicate no clear nonzero slope (most of the data are clustered in a central cloud), and therefore no transformation is necessary.

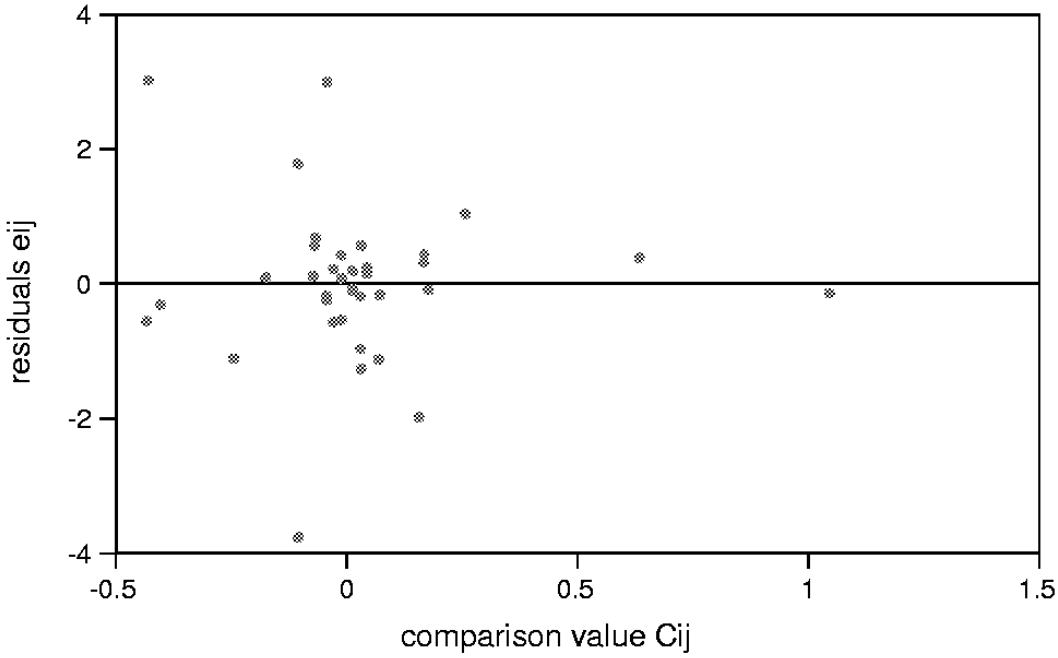


Figure 7.12 Median polish residuals plot showing random scatter around e_{ij}=0

7.3.2 The Friedman Test

The Friedman test is the most common nonparametric test used for the randomized complete block design. It computes the ranks of the data only within each block, not making cross-comparisons between blocks. Treatment effects are determined from the within-block ranks each treatment has received. The Friedman test is an extension of the sign test, and reduces to the sign test when comparing only two treatment groups. Its advantages and disadvantages in comparison to the analysis of variance are the same as that of the sign test to the t-test. When the errors ε_{ij} can be considered normal, the ANOVA should be preferred. For the many situations where the errors are not normal, the Friedman test will generally have equal or greater power to detect differences between treatment groups, and should be performed. The Friedman test is especially useful when the data can be ranked but differences between observations cannot be computed, such as when comparing a <1 to a 5.

7.3.2.1 Null and alternate hypotheses

The Friedman test is used to determine whether

H_0: the median values for k groups of data are identical, or

H_1: at least one median is significantly different.

As with the Kruskal-Wallis test, the test does not provide information on which medians are significantly different from others. That information must come from a multiple comparison test.

7.3.2.2 Computation of the exact test

Rank the data within each block from 1 to k, from smallest to largest. If the null hypothesis is true, the ranks within each block will vary randomly with no consistent pattern. Summing across blocks, the average rank for each treatment group will be similar for all groups, and also be close to the overall average rank. When the alternative hypothesis is true, the ranks in most of the blocks for one or more of the groups will be consistently higher or lower than others. The average group rank for those groups will then differ from the overall average rank. A test statistic Xf is constructed which uses the square of the differences between the average group ranks and the overall rank, to determine if groups differ in magnitude.

The exact test statistic for the Friedman test is a function of both the number of blocks and treatments. Iman and Davenport (1980) state that the exact test should be used for all cases where the number of treatment groups plus the number of blocks (k + n) is ≤ 9. For larger sample sizes a large sample approximation is sufficiently accurate for use. When the number of blocks n is small, the F approximation should be preferred over the chi-square approximation (see the next section).

Should the exact test be required, compute the exact test statistic Xf as shown for the large sample approximation of the following section. Xf is computed identically for both the exact form and large sample approximation. When ties occur, either a corrected large sample

approximation must be used, or the rank transform (F approximation) calculated. The rank transform may be easier to compute.

7.3.2.3 Large sample approximation

For years the Friedman test statistic was approximated using a chi-square distribution with k−1 degrees of freedom. This is the approximation used by statistics packages, and is presented here because of its common use. However, it does not take into account the number of blocks in the data set, and can be in serious error for small n and small α ($\alpha < 0.1$) (Iman and Davenport, 1980). An F approximation which is more accurate for small n is also available. It can be computed from the chi-square approximation, or directly from the data as a rank transform method (an analysis of variance on the within-block ranks R_{ij}).

The box on the next page outlines the computation process for the large sample approximation to the Friedman test statistic.

Example 3, continued.

The Friedman test is used to determine if the median concentration of periphyton mercury differs for the 6 stations along the South River of Virginia. The boxplots of this data were shown in figure 7.10, and the data given in table 7.3. The within-block ranks are given below. For 6 blocks (date) and 6 stations, sample sizes are large enough to employ an approximation, so the preferred F approximation is computed.

Table 7.5 Within-Block Ranks of the Table 7.3 data

Station:	1	2	3	4	5	6
Date						
1	1	4	2	3	5	6
2	1.5	1.5	3	4	6	5
3	1.5	1.5	3	4	6	5
4	2	1	3	4	6	5
5	1	2	3	4	6	5
6	1	2	5	3	6	4
$\bar{R}_j$	1.33	2.0	3.17	3.67	5.83	5.0

overall median = (k+1)/2 = 3.5

The Friedman test

Situation Measurements of k treatment groups are performed on the same or related sets of subjects, called blocks. There are n blocks. One observation is made on each group-block combination ($N = k \bullet n$).

Computation Within each block, observations are ranked from 1 to k, smallest to largest. These within-block ranks R_{ij} are then used to compute the average group rank $\bar{R}_j$ for each of the j=1,k treatment groups:

$$\bar{R}_j = \frac{\sum_{i=1}^{n} R_{ij}}{n} .$$

Test Statistic The average group rank $\bar{R}_j$ is compared to the overall average rank $\bar{R} = (k+1)/2$ in the test statistic Xf:

$$Xf = \frac{12\,n}{k\,(k+1)} \sum_{j=1}^{k} \left[\bar{R}_j - \frac{k+1}{2}\right]^2 .$$

Xf is compared either to an exact table or approximated by a chi-square distribution with (k−1) degrees of freedom. However, a better approximation is available which is compared to an F distribution (Iman and Davenport, 1980). This form is more accurate for small n.

$$f = \frac{(n-1)\,Xf}{n\,(k-1) - Xf} .$$

Tied data When observations are tied within a block, assign the average of their ranks to each. Xf must be corrected when more than a few ties occur.

$$Xf = \frac{12\,n}{k\,(k+1) - \frac{1}{n(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} (t_{ij}\,(j^3-j))} \sum_{j=1}^{k} \left[\bar{R}_j - \frac{k+1}{2}\right]^2 .$$

where t_{ij} equals the number of ties of extent j in row i. The test statistic f is then computed from this corrected Xf as above. An alternative to computing Xf and then f is the rank transform ANOVA (next section).

Decision Rule To reject H_0: the median of every group is identical, versus
H_1: at least one median differs

<u>Exact test:</u> Reject H_0 if $Xf > x_\alpha$, the $(1-\alpha)$th quantile of the Friedman test statistic distribution from table B7 of the Appendix; otherwise do not reject H_0.

<u>F-approximation:</u> Reject H_0 if $f \geq F_{1-\alpha,\,k-1,\,(n-1)(k-1)}$ the $1-\alpha$ quantile of an F distribution with k−1 and (n−1)(k−1) degrees of freedom; otherwise do not reject H_0.

There are only two ties, so ignoring the formula for the tie correction to the variance,

$$Xf = \frac{12(6)}{6(7)}\sum_{j=1}^{6}\left[\bar{R}_j - \frac{7}{2}\right]^2 = \frac{12}{7}\sum\ (-2.17)^2 + (-1.5)^2 + (-0.33)^2 + (0.17)^2 + (2.33)^2 + (1.5)^2.$$

$$= \frac{12}{7} \bullet 14.78$$

$= 25.33$. This can be compared to a chi-square distribution having $k-1 = 5$ df.

To be more exact, the tie correction will be computed. For rows i=1,4,5,6 there are no ties. So for j=1, $t_{ij} = 6$ (there are 6 "ties" of extent 1), and for j=2 to 6, $t_{ij} = 0$ (no true ties). For these four rows

$$\sum_{j=1}^{k} (t_{ij}\ (j^3-j)) = 6(1-1)+0(8-2)+0(27-3)+0(64-4)+0(125-5)+0(216-6) = 0.$$

Rows without ties will always add to zero. Also note that "ties" of extent 1 will always contribute 0 to the sum, as $1^3-1 = 0$. For rows i=2 and 3 there is one pair of tied values per row. Thus for j=1, $t_{ij} = 4$ (4 single values); for j=2, $t_{ij} = 1$ (1 tie of extent 2), and for j=3 to 6, $t_{ij} = 0$ (no triplicates, etc.). For each of these two rows

$$\sum_{j=1}^{k} (t_{ij}\ (j^3-j)) = 4(1-1)+1(8-2)+0(27-3)+0(64-4)+0(125-5)+0(216-6) = 6.$$

Therefore $\sum_{i=1}^{n}\sum_{j=1}^{k} (t_{ij}\ (j^3-j)) = 0+6+6+0+0+0 = 12$, and

$$Xf = \frac{12 \bullet 6}{6\ (7) - \frac{1}{6(5)} \bullet 12} \bullet 14.78 = 25.58$$

which can be compared to a chi-square distribution with 5 degrees of freedom.

The better approximation is the F approximation, or

$$f = \frac{(5)\ 25.58}{6\ (5) - 25.58} = 28.94$$, which is compared to $F_{0.95,\ 5,\ 25} = 4.5$

Therefore reject H_0 that the medians are the same with a p-value of <0.0001.

7.3.2.4 Rank transform approximation: analysis of variance: on within-block ranks

Again an approximation to the exact test statistic may be computed by performing the parametric two-factor ANOVA on the ranks. For the Friedman test, the appropriate ranks are the within-block ranks of table 7.5. Ties are automatically corrected for by assigning the average rank to all ties within a block. A two-factor ANOVA on the within-block ranks has an ANOVA table as in section 7.3.4. The resulting F statistic, the ratio of the MST for the treatment group over the MSE, is the same as the statistic f derived from the chi-square approximation above. Thus the ANOVA on within-block ranks gives a better approximation

than does Xf for the cases ($\alpha < 0.1$ and small n) where the chi-square approximation is inaccurate (Groggel, 1987).

Example 3, continued.
The ANOVA table for the within-block ranks of table 7.5 is:

Source	df	SS	MS	F	p-value
Date (block)	5	0.000	0.000		
Station	5	88.667	17.733	28.93	<0.0001
Error	25	15.333	0.613		
Total	35	104.000			

Note that all differences between blocks have been nullified by transforming the data to the identical within-block ranks, 1 to k. As the blocks all have the same values within them, the block sum of squares equals 0. Also note that the F statistic is identical to that previously calculated from the large-sample approximation after tie correction. Therefore the ANOVA on within-block ranks provides a convenient way to avoid the complicated tie correction to the Friedman statistic.

7.3.3 Median Aligned-Ranks ANOVA

The Friedman test is the multi-treatment equivalent of the sign test. In Chapter 6 the signed-rank test was presented in addition to the sign test, and was favored over the sign test when the differences between the two treatments were symmetric. In this section a multi-treatment equivalent to the signed-rank test is presented, called the Median Aligned-Ranks ANOVA (MARA). MARA is one of several possible extensions of the signed-rank test; others include Quade's test (Conover, 1980). Groggel (1987) and Fawcett and Salter (1984) have shown that an aligned-rank method has substantial advantages in power over other possible signed-rank extensions.

Friedman's test avoids any comparisons across blocks, just as the sign test avoids comparisons of the magnitudes of paired differences across blocks. This avoids the confusion produced by block-to-block differences, but does not take advantage of the information contained in such comparisons. MARA allows comparisons between blocks by first subtracting the within-block median from all of the data within that block. This "aligns" the data across blocks to a common center. It is equivalent to the ranking of block-to-block differences done in the signed-ranks test. To derive the benefits of cross-block comparisons, a cost is incurred. This is an assumption that the residuals ε_{ij} are symmetric. Symmetry can be evaluated by estimating the residuals using median polish, and producing a boxplot as in figure 7.11.

Note that just as for the Friedman's test and two-way ANOVA without replication there are $(k-1)(n-1)$ error degrees of freedom, $(n-1)$ less than a one-way ANOVA. MARA is a two-

factor analysis, with alignment contributing the block effect. However, MARA is computed using a one-way ANOVA on the aligned ranks, so the correct F-test will differ from that performed automatically by a computerized analysis. The error degrees of freedom must be $(k-1)(n-1)$, not $k(n-1)$ as for a one-way ANOVA. MARA is identical to the aligned ranks procedure of Fawcett and Salter (1984), except that the block median is used for alignment rather than the block mean.

The Median Aligned-Ranks ANOVA test

Situation Measurements of k treatment groups are performed on the same or related sets of subjects, called blocks. There are n blocks. One observation is made on each group-block combination ($N = k \bullet n$).

Computation Within each of the n blocks, the observations are aligned by subtracting the block median, forming the aligned o_{ij}.

$o_{ij} = (y_{ij} - b_i)$, where block median $b_i = [\text{median}(y_{ij}), j=1,...k]$

The o_{ij} are then ranked from 1 to N, forming aligned ranks AR_{ij}:

$AR_{ij} = \text{rank}(o_{ij})$.

Test Statistic One-way analysis of variance is computed on the AR_{ij} However, the F statistic is $F = MST/MSE$, where the error degrees of freedom are $(n-1)$ less than in a one-way ANOVA because of the alignment procedure. The ANOVA table is:

Source	df	SS	MS	F
Treatment	$(k-1)$	SST	$SST/(k-1)$	MST/MSE
Error	$(k-1)(n-1)$	SSE	$SSE/[(k-1)(n-1)]$	
Total	$n(k-1)$	Total SS		

Tied data Average ranks are assigned to all tied o_{ij}.

Decision Rule To reject H_0: the median of every group is identical, versus

H_1: at least one median differs

Reject H_0 if $F \geq F_{1-\alpha,\, k-1,\, (n-1)(k-1)}$ the $1-\alpha$ quantile of an F distribution with $k-1$ and $(n-1)(k-1)$ degrees of freedom; otherwise do not reject H_0.

7.3.3.1 Null and alternate hypotheses

The null and alternate hypotheses are identical to those of the Friedman test

H_0: the median values for k groups of data are identical, or

H_1: at least one median is significantly different.

Here, however, it is assumed that the residuals ε_{ij} are symmetric. MARA does not provide information on which medians are significantly different from others. That must come from a multiple comparison test.

7.3.3.2 Computation

MARA is a rank transform approximation test; p-values for an exact test have not been computed.

Example 3, continued

The aligned o_{ij} for the periphyton mercury data were computed during the first step of the median polish, and listed in table 7.4. These o_{ij} are then ranked from 1 to N=36 to form aligned ranks, which are presented in table 7.6:

Table 7.6 Aligned Ranks of the Table 7.4 data

Station:	1	2	3	4	5	6
Date						
1	9	19	14	18	24	30
2	5.5	5.5	12	26	36	31
3	10.5	10.5	15	23	33	28
4	8	7	13	25	34	32
5	3	4	17	20	35	29
6	1	2	22	16	27	21

A one-way analysis of variance is conducted on these aligned ranks. However, the computerized F-test is ignored, as the error degrees of freedom used were n(k−1)=30, and do not reflect the alignment process. The appropriate ANOVA table and F-test are below, and the p-value shows that H_0 is to be rejected. Significant differences are found between treatment group medians:

Source	df	SS	MS	F	p-value
Station	5	3290.3	658.1	27.71	<0.0001
Error	25	593.7	23.8		
Total	30	3884.0			

7.3.4 Parametric Two-Factor ANOVA Without Replication

The traditional parametric test for the randomized complete block design is again an analysis of variance -- a two-factor ANOVA without replication. One factor is the contrast between treatment groups while the second is the block effect. There is one observation (no replicates) per treatment-block combination. The block effect is of no interest except to remove its masking of the treatment effect, so no test for its presence is required.

7.3.4.1 Null and alternate hypotheses

The hypotheses are similar to those of the Friedman and MARA tests, except that treatment group means, rather than medians, are being tested.

H_0: the k treatment group means are identical, $\mu_1 = \mu_2 = ... = \mu_k$, versus

H_1: at least one mean is significantly different.

The ANOVA model is identical to that for all of the tests of this section:

$$y_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij},$$

where y_{ij} is the individual observation in block i and group j;

μ is the overall mean,

α_j is the "jth group effect", j=1,k

β_i is the "ith block effect", i=1,n

ε_{ij} is the residual or "error" between the individual observation and the combined group and block effects.

Here, however, it is assumed that the residuals ε_{ij} follow a normal distribution. ANOVA does not provide information on which means differ from others. That must come from a multiple comparison test.

7.3.4.2 Computation

As with other analysis of variance procedures, the treatment and error mean squares are computed, and their ratio forms the F statistic to be compared to a table of the F distribution for evaluation of its significance. Again there are k treatment groups and n blocks.

In comparison to a one-way ANOVA without blocking, the error sum of squares SSE is split into two parts, the SSE and the sum of squares for the block effect SSB. The variation due to differences between blocks is thereby removed from the background noise (MSE). If there is an appreciable block effect, removal of the SSB lowers the SSE and MSE in comparison to their values for a one-way ANOVA. This produces a higher F statistic, allowing the treatment effect to be more easily discerned.

Example 3, continued

An analysis of variance is calculated directly on the periphyton mercury data. The ANOVA table is:

Source	df	SS	MS	F	p-value
Date	5	3.26	0.65		
Station	5	230.13	46.03	26.15	<0.0001
Error	25	44.02	1.76		
Total	35	277.40			

The null hypothesis is again soundly rejected. The treatment group means are declared different at any reasonable alpha level. As in all of the tests applied to this data set, the block effect (Date) is minimal.

Two-factor ANOVA without replication

Situation Measurements of k treatment groups are performed on the same or related sets of subjects, called blocks. There are n blocks. One observation is made on each group-block combination (N = k•n).

Computation Sums of squares for treatment, block and error are computed using the following formula. These are divided by their appropriate degrees of freedom to form mean squares.

$$\text{SST} = \frac{\sum^{k}\left[\sum^{n} y\right]^2}{n} - \frac{\left[\sum^{k}\sum^{n} y\right]^2}{kn} \qquad \text{SSB} = \frac{\sum^{n}\left[\sum^{k} y\right]^2}{k} - \frac{\left[\sum^{k}\sum^{n} y\right]^2}{kn}$$

$$\text{SSE} = \text{Total SS} - \text{SST} - \text{SSB} \qquad \text{Total SS} = \sum^{k}\sum^{n} y^2 - \frac{\left[\sum^{k}\sum^{n} y\right]^2}{kn}$$

Test Statistic The F statistic is computed as F = MST/MSE,. The ANOVA table is:

Source	df	SS	MS	F	p-value
Treatment	k−1	SST	SST/(k−1)	MST/MSE	
Block	n−1	SSB	SSB/(n−1)		
Error	(k−1)(n−1)	SSE	SSE/[(k−1)(n−1)]		
Total	N−1	Total SS			

Tied data No corrections necessary.

Decision Rule To reject H_0: the mean of every group is identical, versus
H_1: at least one mean differs

Reject H_0 if $F \geq F_{1-\alpha,\, k-1,\, (n-1)(k-1)}$ the $1-\alpha$ quantile of an F distribution with k−1 and (n−1)(k−1) degrees of freedom; otherwise do not reject H_0.

7.4 Multiple Comparison Tests

In most cases the analyst is interested not only in whether group medians or means differ, but which differ from others. This is information not supplied by the tests presented in the previous sections, but by methods called multiple comparison tests (MCTs). MCTs compare all possible pairs of treatment group medians or means, and are performed only after the null hypothesis of

"all medians or means identical" has been rejected. Of interest is the "pattern" of group medians or means:

$$\text{group A} \cong \text{group B} << \text{group C},$$

etc. MCT's are not efficient methods for contrasting specific sets of groups known to be of interest before an ANOVA or Kruskal-Wallis test is done, such as a treatment versus a control. Other tests are available for making specific contrasts. Instead, MCTs compare all possible combinations of treatment group centers, ranking the centers in order and indicating which are similar or different from others.

Stoline (1981) reviews the many types of parametric multiple comparison tests. Campbell and Skillings (1985) discuss nonparametric multiple comparisons.

7.4.1 Parametric Multiple Comparisons

Parametric MCTs compare treatment group means. They often calculate a "least significant range" or LSR, the distance between any two means which must be exceeded in order for the two groups to be considered significantly different at a significance level α.

$$\text{If } \left|\bar{y}_1 - \bar{y}_2\right| > LSR = R\sqrt{s^2/n}\,, \qquad \bar{y}_1 \text{ and } \bar{y}_2 \text{ are significantly different.}$$

The statistic R is analogous to the t-statistic in a t-test. R depends on the test used (is some function of either a t- or studentized range statistic q), the error degrees of freedom from the ANOVA, and on α. The variance s^2 is just the MSE from the ANOVA. Parametric MCTs can be classified into four types, based on their method of computation and on whether a pairwise or overall α level is used (figure 7.13).

	α pairwise	**α overall**
MST (equal n only)	Duncans Multiple Range test SNK	REGWQ * REGWF *
SIM (equal or unequal n)	Fisher's t-tests (LSD)	Tukey * Scheffe Bonferroni

Figure 7.13 Types of Parametric Multiple Comparison Tests

Methods with an asterisk * in figure 7.13 have the most power to detect differences between group means of those methods using the overall error rate. The REGW methods are the most powerful (have the smallest LSR) for equal sample sizes, though Tukey's test is close in power. For unequal sample sizes, Tukey's method is the most powerful of those listed. Therefore

Tukey's method is a generally applicable and powerful multiple comparison test for a variety of situations.

Multiple-stage tests, MST, are valid only when group sample sizes are equal. Examples are the Duncan's, Student-Newman-Keuls (SNK), and REGW tests. Their R statistic varies for each pairwise comparison as a function of the number of group means in between the two being compared. A new least significant range $(\bar{y}_1 - \bar{y}_2)$ must then be computed for each pairwise comparison of means. If sample sizes were unequal, test results could be non-intuitive, as in: A > B, B > C, but A = C where "A > B" means that A is larger and significantly different from B, and "A = C" means A is not significantly different from C. This could arise if B had a large sample size so that comparisons involving it had a lower LSR than those not involving B. Thus MSTs are valid only for equal sample sizes within all groups.

Simultaneous inference methods, SIM, are valid for both equal and unequal group sample sizes. Examples are Tukey's, Sheffe's, and Fisher's t-tests. These tests use one R value to calculate a single least significant range for all pairwise comparisons. The harmonic mean

$$\text{harmonic mean of } n_1 \text{ and } n_2 = \frac{2\, n_1\, n_2}{n_1 + n_2}$$

is substituted for n in the case of unequal group sample sizes. So for unequal sample sizes a SIM should be used.

The second classification criteria for MCTs is based on the type of error rate α used for comparisons (figure 7.13). One class of tests uses the stated α level for each pairwise comparison (α_p= pairwise error rate). When there are multiple comparisons each having a pairwise error rate of α, the overall probability of declaring at least one false difference (the overall error rate α_o) is much greater than α_p. This overall error rate is the error rate for the "pattern" of group means, and is more often of interest than a pairwise error rate in water resources applications. For example, when comparing six group means, there are (6•5)/2 = 15 pairwise comparisons. If $\alpha_p = 0.05$ is used for each test, then there will be an overall error rate $\alpha_o = 1 - (1 - \alpha_p)^{15} = 0.54$ of making at least one error in the overall comparisons of the six group means.

Unfortunately, the distinction between the overall and pairwise error rates is often not understood, and pairwise rates are presented as if they were overall rates. The pairwise rate is much like the probability of being robbed today, while the overall rate is like the probability of ever being robbed in your lifetime. To claim that the (very small) probability of being robbed today is actually the probability of ever being robbed leads to a false sense of security. Similarly, citing that according to a Duncan's multiple range test, A > B = C = D > E = F with an error rate of $\alpha = 0.05$ when in fact 0.05 was used for each test, also presents a false sense of security in the results.

Duncan's test is often used in this incorrect fashion. Individual paired differences found at the $\alpha = 0.05$ level results in the overall rate of at least one error in the pattern of group means at something higher, such as the 0.54 chance for the 6 groups above. When the primary interest is in the overall pattern and its accuracy, methods which set the error rate equal to the overall α, such as Tukey's test, should be performed.

Some authors report only results of a MCT, usually Duncan's multiple range test, skipping the required prior ANOVA F-tests. **NEVER DO THIS!** The likely reason that this has been done is that ANOVA did not find significant differences at a true (overall) significance level of 0.05, but the Duncan's test did find differences. Why does this occur? Duncan's test was performed at a pairwise significance level of 0.05, but at an overall level of something much higher (0.54 for the six means above). An overall error level of 0.54 states there a 54 percent chance that two means will be declared significantly different when in fact they are not. An ANOVA at $\alpha = 0.54$ would also be "significant" (the p-value is somewhere below 0.54), but a test having this large an error rate is essentially useless! ANOVA should always be performed first as the appropriate test for determining whether any differences occur between group means. If they do not, stop there. By performing only a MCT, an α=0.54 test is conducted while declaring it to be an α = 0.05 test of whether differences occur. This is quite misleading.

7.4.1.1 Assumptions

All MCTs discussed thus far have the same assumptions as does ANOVA -- data within each treatment group are normally distributed, and each treatment group has equal variance. Violations of these assumptions will result in a loss of power to detect differences which are actually present.

7.4.1.2 Computation of Tukey's test

Two group means $\bar{y}_i$ and $\bar{y}_j$ can be considered different if

$$\left|\bar{y}_i - \bar{y}_j\right| > q_{(1-\alpha),\, k,\, N-k} \bullet \sqrt{MSE / n}$$

where
- q is the studentized range statistic from Neter, Wasserman and Kutner (1985),
- α is the overall significance level,
- k is the number of treatment group means compared,
- N−k are the degrees of freedom for the MSE, and
- n is the sample size per group.

For unequal sample sizes

$$\left|\bar{y}_i - \bar{y}_j\right| > q_{(1-\alpha),\, k,\, N-k} \bullet \sqrt{MSE \bullet \frac{n_i + n_j}{2\, n_i\, n_j}}$$

where n has been replaced with the harmonic mean of the unequal sample sizes for the two groups being compared, n_i and n_j. For only two groups, q becomes the student's t statistic, and

Tukey's test is identical to Fisher's all-possible t-tests. Formulas for other MCTs can be found in SAS Institute (1985).

Example 4

Knopman (1990) tested wells located in the Appalachian mountains of Pennsylvania to see if their specific capacities differed among four rock types -- dolomites, limestones, siliciclastics (sandstones, shales, etc.), and metamorphic plus igneous rocks. To make the data more nearly normal, the natural log of specific capacity was used. A subset of 200 observations across the four rock types were randomly selected from the over 4000 original observations. This subset is presented in Appendix C7. Boxplots are shown in figure 7.14. The ANOVA table below indicates that the log specific capacities differed significantly between the four rock types.

Source	df	SS	MS	F	p-value
Rock type	3	54.03	18.010	4.19	0.007
Error	196	842.15	4.297		
Total	199	896.18			

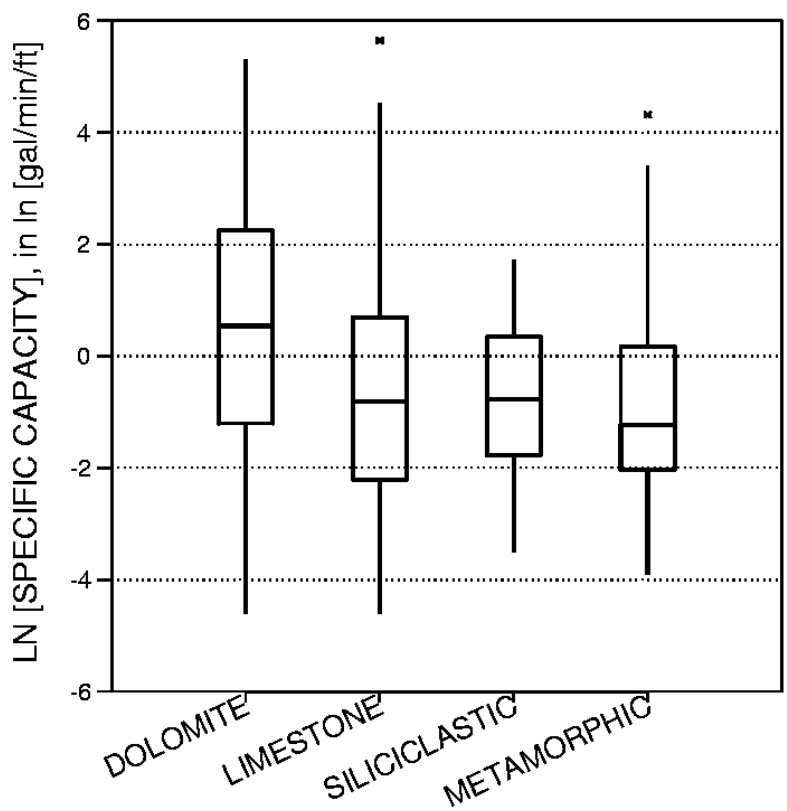


Figure 7.14 Natural logs of specific capacity of wells in four rock types, Pennsylvania

Since the null hypothesis is rejected, Tukey's test can be computed. The four group means are :

$\bar{y}$ [dolomite] $= \bar{y}_d = 0.408$ $\bar{y}$ [limestone] $= \bar{y}_l = -0.688$

$\bar{y}$ [siliciclastic] $= \bar{y}_s = -0.758$ $\bar{y}$ [metamorphic] $= \bar{y}_m = -0.894$

The least significant range LSR is computed as:

$$\text{LSR} = q_{(0.95,\,4,\,196)} \bullet \sqrt{4.297/50} \cong q_{(0.95,\,4,\,\infty)} \bullet \sqrt{4.297\ /\ 50} = 3.63 \bullet 0.293$$
$$= 1.06$$

Therefore, any group means of log specific capacity which differ by more than 1.06 are significantly different by the Tukey's multiple comparison test. $\bar{y}_d$ is then seen to be significantly different and larger than the other three groups, which are not significantly different from each other, or:

$$\bar{y}_d > \bar{y}_l = \bar{y}_s = \bar{y}_m$$

REGWQ could also be computed because sample sizes in each subset group are equal. The choice of REGWQ versus Tukey's would largely depend on which were available. First the k group means are ordered by magnitude ($\bar{y}_d$, $\bar{y}_l$, $\bar{y}_s$, $\bar{y}_m$). The first comparison is made between the extremes, $\bar{y}_d$ versus $\bar{y}_m$. The studentized range is again used, accounting for the number of means between and including the two being compared; k=4 in this first case. If this test proves to be significant, the two possible comparisons with p =k–1 intervening group means are made -- $\bar{y}_d$ versus $\bar{y}_s$ and $\bar{y}_l$ versus $\bar{y}_m$. Continue working inward until an insignificant difference is found. No comparisons of group means contained between means already found to be insignificant need be made.

For REGWQ, two group means differ at an overall significance level α if :

$$\bar{y}_i - \bar{y}_j > q_{\alpha_p,\,p,\,N-p} \bullet \sqrt{\text{MSE}\ /\ n}$$

$$\text{where } \alpha_p = 1 - (1-\alpha)^{p/k} \quad \text{for } p < (k-1)$$
$$= \alpha \quad \text{for } p \geq (k-1).$$

Using the log specific capacity data, comparing $\bar{y}_d$ versus $\bar{y}_m$ using $\alpha_p = \alpha = 0.05$: the least significant range = $q_{0.05,\,4,\,196} \bullet \sqrt{4.297\ /\ 50} = 1.06$, identical to Tukey's LSR. Therefore $\bar{y}_d > \bar{y}_m$. Next, compare $\bar{y}_d$ versus $\bar{y}_s$ and $\bar{y}_l$ versus $\bar{y}_m$. Both of these have p=3 and an LSR of $q_{0.05,\,3,\,197} \bullet \sqrt{4.297\ /\ 50} = 3.31 \bullet 0.293 = 0.97$. Therefore $\bar{y}_d > \bar{y}_s$ and $\bar{y}_l = \bar{y}_m$. Since the limestone and metamorphic group means are not significantly different there is no reason to test the siliciclastic versus the metamorphic group means. For the final comparison, $\bar{y}_d$ is compared to $\bar{y}_l$. The LSR is based on p=2 and $\alpha_p = 1 - (0.95)^{2/4} = 0.025$. Therefore LSR = $q_{0.025,\,2,\,198} \bullet \sqrt{4.297\ /\ 50} = 3.31 \bullet 0.293 =$ 0.97. So $\bar{y}_d > \bar{y}_l$ and the overall pattern is again:

$$\bar{y}_d > \bar{y}_l = \bar{y}_s = \bar{y}_m$$

7.4.2 Nonparametric Multiple Comparisons

Statisticians are actively working in this area (see Campbell and Skillings, 1985). The simplest procedures for performing nonparametric multiple comparisons are rank transformation tests. Ranks are substituted for the original data, and a multiple comparison test such as Tukey's is

performed on the ranks. These are logical follow-ups to the rank transform approximation approaches to the Kruskal-Wallis, Friedman, and two-way ANOVA tests previously presented.

For the one-way situation, Campbell and Skillings (1985) recommend a multiple-stage test using the Kruskal-Wallis (KW) statistic. The process resembles the REGWQ test above. After a significant KW test occurs for k groups, place the groups in order of ascending average rank. Perform new KW tests for the two possible comparisons between $p = (k-1)$ groups, noting that this involves re-ranking the observations each time. If significant results occur for one or both of these tests, continue attempting to find differences in smaller subsets of $p < (k-1)$. In order to control the overall error rate, follow the pattern of REGWQ for the critical alpha values:

$$\alpha_p = 1 - (1-\alpha)^{p/k} \quad \text{for } p < (k-1)$$
$$= \alpha \quad \text{for } p \geq (k-1)$$

Example 4 continued

First, Tukey's test will be performed on the ranks of the Pennsylvania log specific capacity data. Then a second nonparametric MCT, the multiple-stage Kruskal-Wallis (MSKW) test using REGWQ alpha levels, is performed.

The ANOVA table for testing data ranks shows a strong rejection of H_0:

Source	df	SS	MS	F	p-value
Rock type	3	38665	12888	4.02	0.008
Error	196	627851	3203		
Total	199	666515			

The four group mean ranks are :

$\overline{R}$ [dolomite] $= \overline{R}_d = 124.11$ $\quad$ $\overline{R}$ [limestone] $= \overline{R}_l = 94.67$

$\overline{R}$ [siliciclastic] $= \overline{R}_s = 95.06$ $\quad$ $\overline{R}$ [metamorphic] $= \overline{R}_m = 88.16$

The least significant range LSR for a Tukey's test on data ranks is computed as:

$$\text{LSR} = q_{(0.95,\,4,\,196)} \bullet \sqrt{3203/50} \cong q_{(0.95,\,4,\,\infty)} \bullet \sqrt{3203/50} = 3.63 \bullet 8.00$$
$$= 29.06$$

Pairs of group mean ranks which are at least 29.06 units apart are significantly different. Therefore (within 0.01) $\overline{R}_d > \overline{R}_s = \overline{R}_l = \overline{R}_m$.

To compute the MSKW test, the first step is merely the Kruskal-Wallis test on the four groups. The overall mean rank $\overline{R}$ equals 100.5. Then

$K=11.54$ $\quad \chi^2_{0.95,(3)} = 7.815$ $\quad p=0.009$ so, reject equality of group medians.

Proceeding, new Kruskal-Wallis tests are performed between the two sets of three contiguous treatment groups: $\overline{R}_d$ vs. $\overline{R}_l$ vs. $\overline{R}_s$ and $\overline{R}_l$ vs. $\overline{R}_s$ vs. $\overline{R}_m$. This requires that the data all be re-ranked each time. Their respective test statistics are denoted K_{dls} and K_{lsm}. The significance level is as in REGWQ, so for (k−1) = 3 groups, $\alpha_p = \alpha = 0.05$.

$K_{dls} = 8.95$ $\chi^2_{0.95,(2)} = 5.99$ p=0.012 so, reject equality of group medians.

$K_{lsm} = 0.61$ p=0.74 group medians not significantly different.

Finally, the k−2 = 2 group comparisons are performed. There is no need to do these for the limestone versus siliciclastic and siliciclastic versus metamorphic comparisons, as the 3-group Kruskal-Wallis test found no differences among those group medians. Therefore the only remaining 2-group comparison is for dolomite versus limestone. The 2-group Kruskal-Wallis test is performed at a significance level of

$\alpha_p = 1 - (0.95)^{2/4} = 0.025$.

$K_{dl} = 5.30$ $\chi^2_{0.975,(1)} = 5.02$ p=0.021 so, reject equality of group medians.

The pattern is the same as for the other MCT's,

$$\text{median}_d > \text{median}_l = \text{median}_s = \text{median}_m.$$

7.5 Presentation of Results

Following the execution of the tests in this chapter, results should be protrayed in an easily-understandable manner. This is best done with figures. A good figure provides a visual confirmation of the outcome of the hypothesis test. Differences between groups are clearly portrayed. A poor figure gives the impression that the analyst has something to hide, and is hiding it effectively! The following sections provide a quick survey of good and bad figures for illustrating differences between three or more treatment groups.

7.5.1 Graphical Comparisons of Several Independent Groups

Perhaps the most common method used to report comparisons between groups is a table, and not a graph. Table 7.7a is the most common type of table in water resources, one which presents only the mean and standard deviations. As has been shown several times, the mean and standard deviation alone do not capture much of the important information necessary to compare groups, especially when the data are skewed. Table 7.7b provides much more information -- important percentiles such as the quartiles are listed as well.

Table 7.7a A simplistic table comparing the four groups of log specific capacity data

	Mean	Std.Dev.
Dolomite	0.408	2.557
Limestone	-0.688	2.360

Siliciclastics	-0.758	1.407
Metamorphic	-0.894	1.761

Table 7.7b A more complete table for the log specific capacity data

	N	Mean	Median	Std.Dev.	Min	Max	P25	P75
Dolomite	50	0.408	0.542	2.557	-4.605	5.298	-1.332	2.264
Limestone	50	-0.688	-0.805	2.360	-4.605	5.649	-2.231	0.728
Siliciclastics	50	-0.758	-0.777	1.407	-3.507	1.723	-1.787	0.381
Metamorphic	50	-0.894	-1.222	1.761	-3.912	4.317	-2.060	0.178

However, neither table provides quick intuitive insight into the data structure. Neither sufficiently illustrates the differences between groups found by the hypothesis tests in example 4, or how they differ.

Histograms are commonly used to display the distribution of one or more data sets, and have been employed to attempt to illustrate differences between three or more groups of data. They are not usually successful. The many crossing lines, coupled with an artificial division of the data into categories, results in a cluttered and confusing graph. Figure 7.15 displays four overlapping histograms, one for each of the data groups. It is impossible to discern anything about the relative characteristics of any of the data groups from this figure. Overlapping histograms should be avoided unless one is purposefully trying to confuse the audience! In figure 7.16, side-by-side bar charts display the same information. This too is confusing and difficult to interpret. From the graph one could not easily say which group had the highest mean or median, much less anything about the groups' variability or skewness. Many business software packages allow speedy production of such useless graphs as these.

Figure 7.17 shows a quantile plot of the same four data groups. The quantile plot far exceeds the histogram and bar chart in clarity and information content. The dolomite group stands apart from the other three throughout most of its distribution, illustrating both the ANOVA and multiple comparison test results. An experienced analyst can look for differences in variability and skewness by looking at the slope and shapes of each group's line. A probability plot of the four groups would have much the same content, with the additional ability to look for departures from a straight line as a visual clue for non-normality.

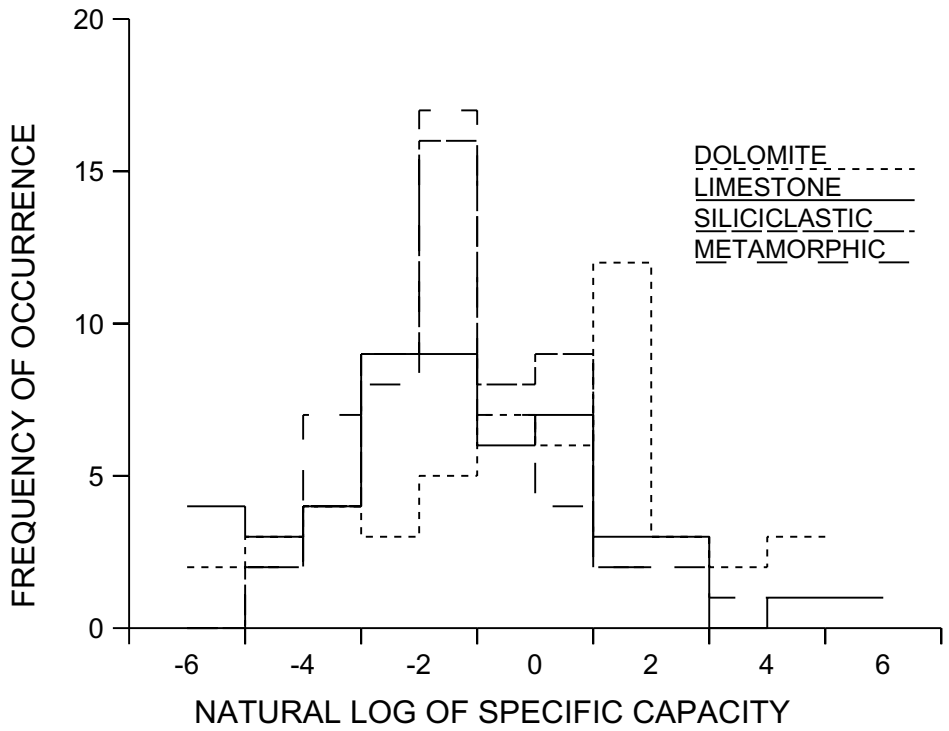


Figure 7.15 Overlapping histograms fail to differentiate between four groups of data

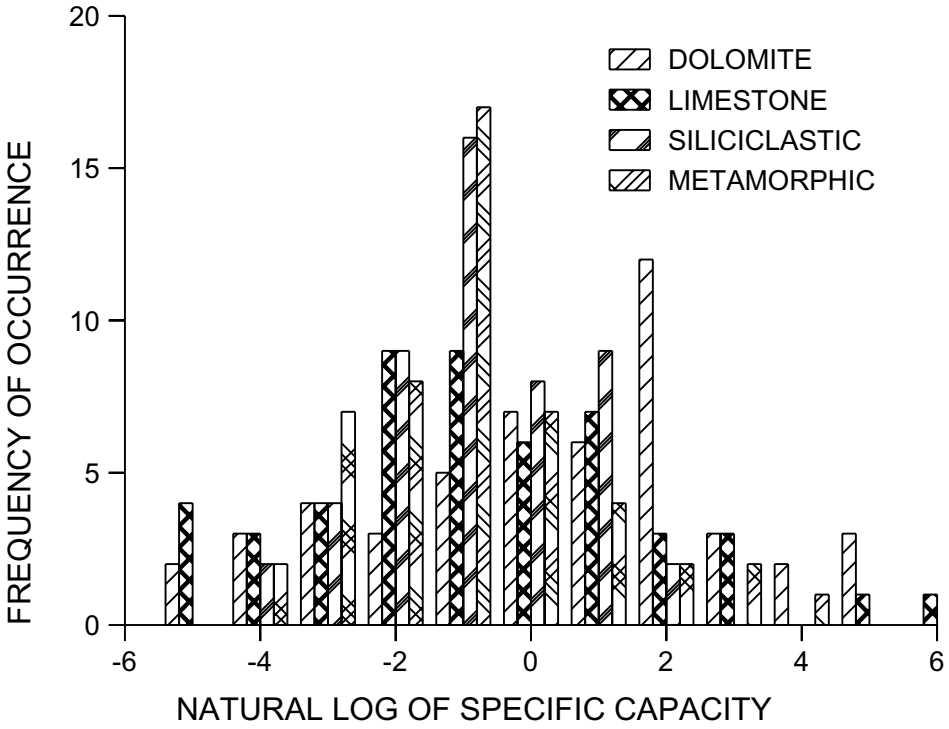


Figure 7.16 Side-by-side bars fail to clearly differentiate between four groups of data

Compare figures 7.15 to 7.17 with boxplots of the log specific capacity data shown previously in figure 7.14. Boxplots clearly demonstrate the difference between the dolomite and other group medians. Variability is also documented by the box height, and skewness by the heights of the top and bottom box halves. See Chapter 2 for more detail on boxplots. Boxplots illustrate the results of the tests of this chapter more clearly than commonly-used alternate methods such as histograms.

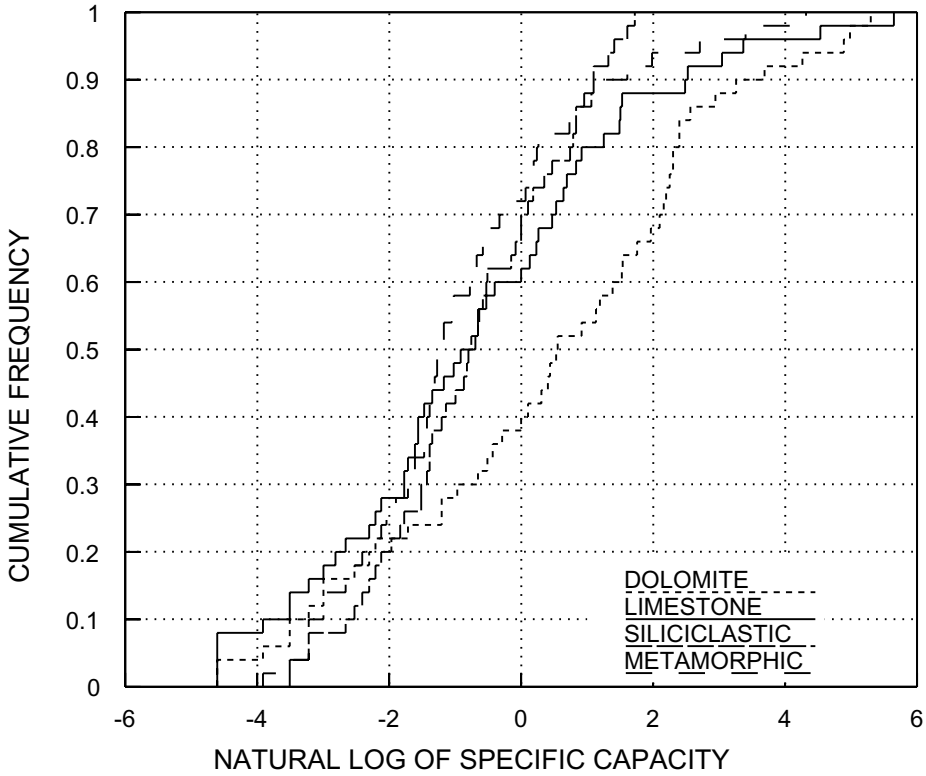


Figure 7.17 Quantile plots differentiate between four groups of data

7.5.2 Presentation of Multiple Comparison Tests

Suppose a multiple comparison test resulted in the following:

$\bar{y}_1 = \bar{y}_2$	$\bar{y}_1 \neq \bar{y}_3$	$\bar{y}_1 \neq \bar{y}_4$	(= : not significantly different)
$\bar{y}_2 = \bar{y}_3$	$\bar{y}_2 \neq \bar{y}_4$		(≠ : significantly different)
$\bar{y}_3 = \bar{y}_4$			

for four treatment groups having $\bar{y}_1 > \bar{y}_2 > \bar{y}_3 > \bar{y}_4$.

The results are often presented in one of the two following formats:

1. Letters

$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
A	AB	BC	C

Treatment group means are ordered, and those having the same letter underneath them are not significantly different. The convenience of this presentation format is that letters can easily be

positioned somewhere within side-by-side boxplots, illustrating the results of a MCT as well as the overall test for equality of all means or medians (see figure 7.18).

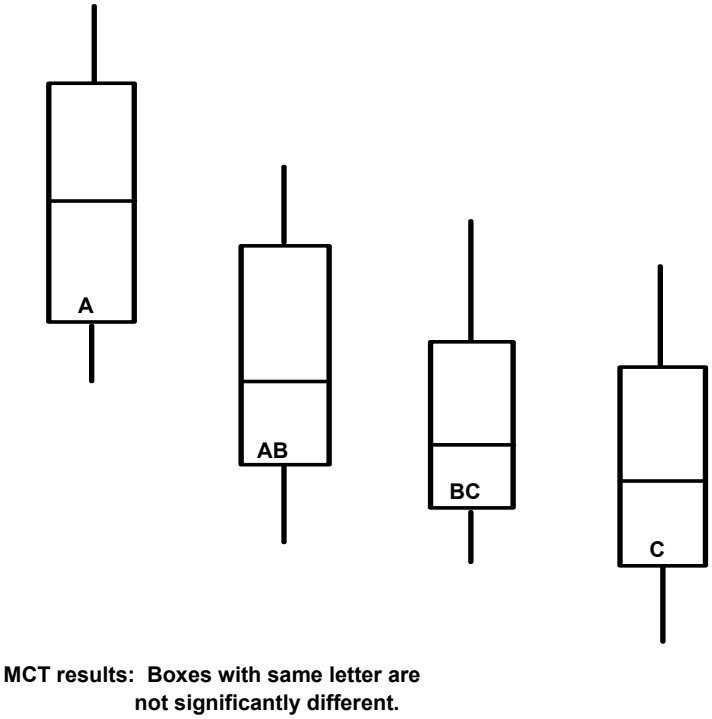


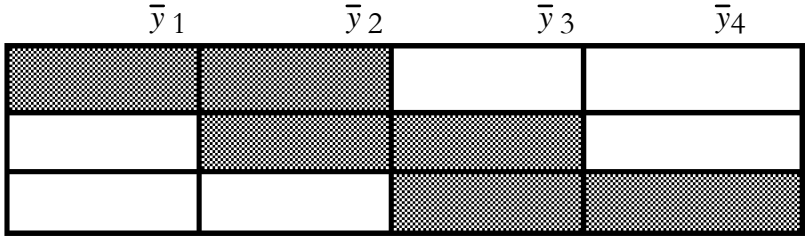
Figure 7.18 Boxplots with letters showing the results of a MCT.

2. Lines

In this presentation format, group means connected by a single unbroken line are not significantly different. This format is suited for inclusion in a table listing group means or medians.

A third method is somewhat more visual:

3. Shaded Boxes



These shaded boxes can be thought of as thick versions of the lines presented above. Group means with boxes shaded along the same row are not significantly different. Shaded boxes allow group means to be ordered by something other than mean or median value. For example, the order of stations going upstream to downstream might be 3,1,2,4. Boxes put in that order show a significant increase in concentration between 3 and 1 and a significant drop off again between 2 and 4. So in addition to displaying multiple comparison test results, the shaded boxes below also illustrate the pattern of concentration levels of the data.

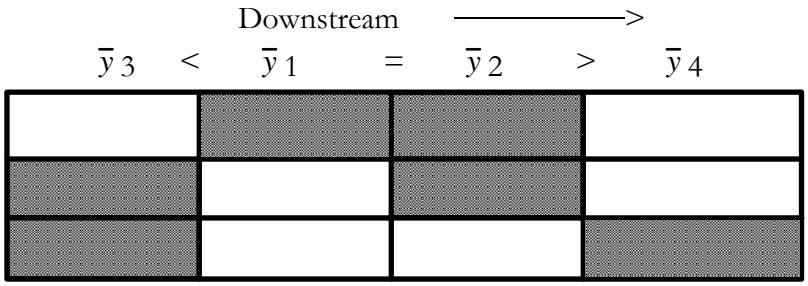


Figure 7.19 Shaded boxes for illustration of a multiple comparison test. Station means not significantly different have boxes shaded within the same row.

Exercises

7.1 Discharge from pulp liquor waste may have contaminated shallow groundwater with caustic, high pH effluent (Robertson, et al., 1984). Determine whether the pH of samples taken from three sets of piezometers are all identical -- one piezometer group is known to be uncontaminated. If not, which groups are different from others? Which are contaminated?

	pH of samples taken from piezometer groups					
BP-1	7.0	7.2	7.5	7.7	8.7	7.8
BP-2	6.3	6.9	7.0	6.4	6.8	6.7
BP-9	8.4	7.6	7.5	7.4	9.3	9.0

7.2 In addition to the waters from granitic terrain given in Exercise 2.3, Feth et al. (1964) measured chloride concentrations of ephemeral springs. These additional data are listed below (use the zero value as is). Test whether concentrations in the three groups are all identical. If not, which differ from others?

	Chloride concentration, in mg/L					
Ephemeral Springs	0.0	0.9	0.1	0.1	0.5	0.2
	0.3	0.2	0.1	2.0	1.8	0.1
	0.6	0.2	0.4			

7.3 The number of Corbicula (bottom fauna) per square meter for a site on the Tennessee River was presented by Jensen (1973). The data are found in Appendix C8. Perform a median polish for the data of strata 1. Graph the polished estimates of year and seasonal effects. Is any transformation suggested by the residuals?

7.4 Test the Corbicula data of strata 1 to determine whether season and year are significant determinants of the number of organisms.

7.5 Test for significant differences in the density of Corbicula between seasons and strata for the 1969 data.

Chapter 8

Correlation

Concentrations of atrazine and nitrate in shallow groundwaters are measured in wells over a several county area. For each sample, the concentration of one is plotted versus the concentration of the other. As atrazine concentrations increase, so do nitrate. How might the strength of this association be measured and summarized?

Streams draining the Sierra Nevada mountains in California usually receive less precipitation in November than in other months. Has the amount of November precipitation significantly changed over the last 70 years, showing a gradual change in the climate of the area? How might this be tested?

The above situations require a measure of the strength of association between two continuous variables, such as between two chemical concentrations, or between amount of precipitation and time. How do they co-vary? One class of measures are called correlation coefficients, three of which are discussed in this chapter. Also discussed is how the significance of that association can be tested for, to determine whether the observed pattern differs from what is expected due entirely to chance. For measurements of correlation between grouped (non-continuous) variables, see Chapter 14.

Whenever a correlation coefficient is calculated, the data should be plotted on a scatterplot. No single numerical measure can substitute for the visual insight gained from a plot. Many different patterns can produce the same correlation coefficient, and similar strengths of relationships can produce differing coefficients, depending on the curvature of the relationship. In Chapter 2, figure 2.1 presented eight plots all with a linear correlation coefficient of 0.70. Yet the data were radically different! Never compute correlation coefficients and assume the data look like those in h of figure 2.1.

8.1 Characteristics of Correlation Coefficients

Correlation coefficients measure of the strength of association between two continuous variables. Of interest is whether one variable generally increases as the second increases, whether it decreases as the second increases, or whether their patterns of variation are totally unrelated. Correlation measures observed co-variation. It does not provide evidence for causal relationship between the two variables. One may cause the other, as precipitation causes runoff. They may also be correlated because both share the same cause, such as two solutes measured at a variety of times or a variety of locations. (Both are caused by variations in the source of the water). Evidence for causation must come from outside the statistical analysis -- from the knowledge of the processes involved.

Measures of correlation (here designated in general as ρ) have the characteristic of being dimensionless and scaled to lie in the range $-1 \leq \rho \leq 1$. When there is no correlation between two variables, $\rho = 0$. When one variable increases as the second increases, ρ is positive. When they vary in opposite directions, ρ is negative. The significance of the correlation is evaluated using a hypothesis test:

$$H_0: \rho = 0 \text{ versus } H_1: \rho \neq 0.$$

When one variable is a measure of time or location, correlation becomes a test for temporal or spatial trend.

8.1.1 Monotonic Versus Linear Correlation

Data may be correlated in either a linear or nonlinear fashion. When y generally increases or decreases as x increases, the two variables are said to possess a monotonic correlation. This correlation may be nonlinear, with exponential patterns, piecewise linear patterns, or patterns similar to power functions when both variables are non-negative. Figure 8.1 illustrates a non-linear monotonic association between two variables -- as x increases, y generally increases by an ever-increasing rate. This nonlinearity is evidence that a measure of linear correlation would be inappropriate. The strength of a linear measure will be diluted by nonlinearity, resulting in a lower correlation coefficient and less significance than a linear relationship having the same amount of scatter.

Three measures of correlation are in common use -- Kendall's tau, Spearman's rho, and Pearson's r. The first two are based on ranks, and measure all monotonic relationships such as that in figure 8.1. They are also resistant to effects of outliers. The more commonly-used Pearson's r is a measure of linear correlation (figure 8.2), one specific type of monotonic correlation. None of the measures will detect nonmonotonic relationships, where the pattern doubles back on itself, like that in figure 8.3.

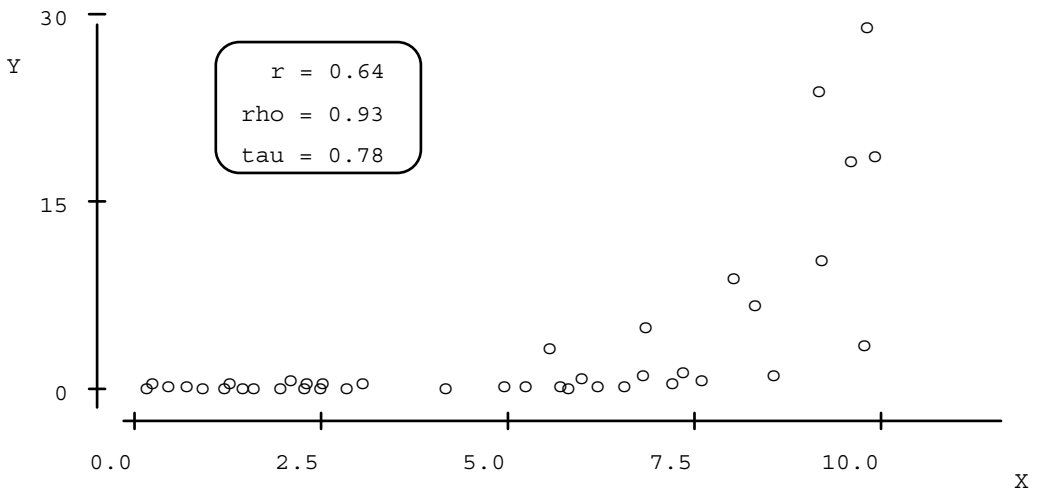


Figure 8.1 Monotonic (nonlinear) correlation between x and y.

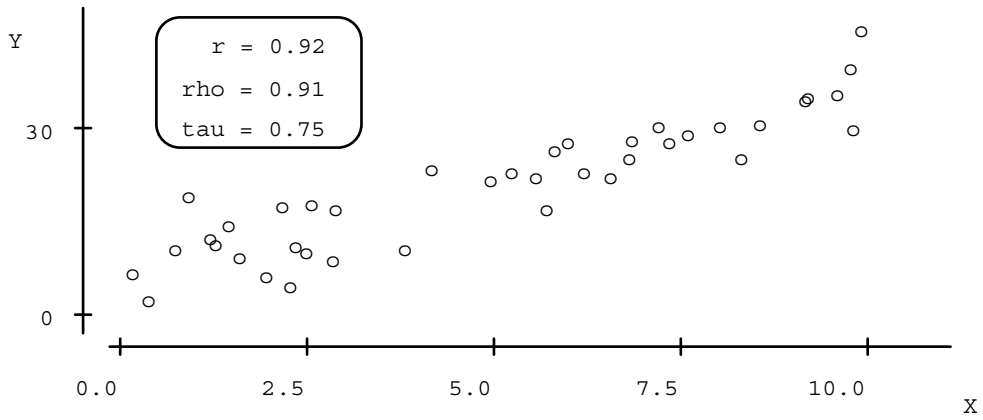


Figure 8.2 Linear correlation between X and Y.

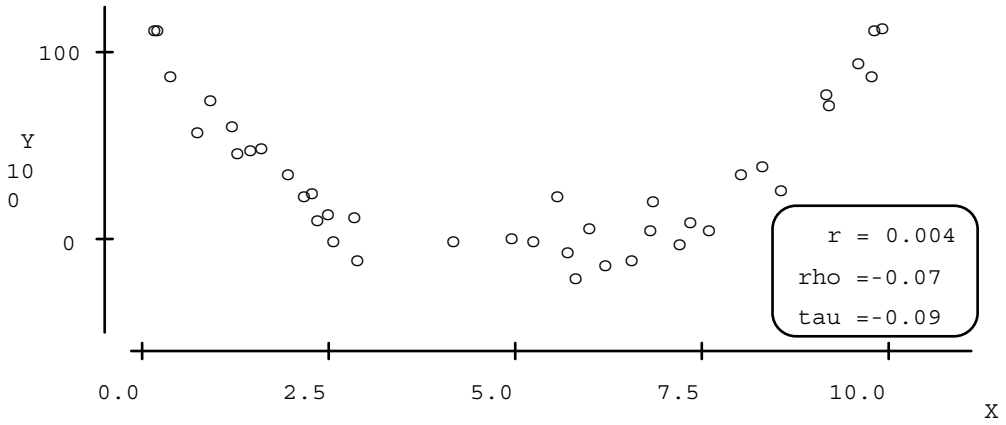


Figure 8.3 Non-monotonic relationship between X and Y.

8.2 Kendall's Tau

Tau (Kendall, 1938 and Kendall, 1975) measures the strength of the monotonic relationship between x and y. Tau is a rank-based procedure and is therefore resistant to the effect of a small number of unusual values. It is well-suited for variables which exhibit skewness around the general relationship.

Because tau (τ) depends only on the ranks of the data and not the values themselves, it can be implemented even in cases where some of the data are censored, such as concentrations known only as less than the reporting limit. This is an important feature of the test for applications to water resources. See Chapter 13 for more detail on analysis of censored data.

Tau will generally be lower than values of the traditional correlation coefficient r for linear associations of the same strength (figure 8.2). "Strong" linear correlations of 0.9 or above correspond to tau values of about 0.7 or above. These lower values do not mean that tau is less sensitive than r, but simply that a different scale of correlation is being used. Tau is easy to compute by hand, resistant to outliers, and measures all monotonic correlations (linear and nonlinear). Its large sample approximation produces p-values very near exact values, even for small sample sizes. As it is a rank correlation method, tau is invariant to monotonic power transformations of one or both variables. For example, τ for the correlation of log(y) versus log(x) will be identical to that of y versus log(x), and of y versus x.

8.2.1 Computation

Tau is most easily computed by first ordering all data pairs by increasing x. If a positive correlation exists, the y's will increase more often than decrease as x increases. For a negative correlation, the y's will decrease more often than increase. If no correlation exists, the y's will increase and decrease about the same number of times.

A two-sided test for correlation will evaluate the following equivalent statements for the null hypothesis H_0, as compared to the alternate hypothesis H_1:

H_0:
- a) no correlation exists between x and y ($\tau = 0$), or
- b) x and y are independent, or
- c) the distribution of y does not depend on x, or
- d) Prob ($y_i < y_j$ for $i < j$) = 1/2.

H_1:
- a) x and y are correlated ($\tau \neq 0$), or
- b) x and y are dependent, or
- c) the distribution of y (percentiles, etc.) depends on x, or
- d) Prob ($y_i < y_j$ for $i < j$) $\neq$ 1/2.

The test statistic S measures the monotonic dependence of y on x. Kendall's S is calculated by subtracting the number of "discordant pairs" M, the number of (x,y) pairs where y decreases as x increases, from the number of "concordant pairs" P, the number of (x,y) pairs where y increases with increasing x:

$$\boxed{S = P - M} \qquad [8.1]$$

where P = "number of pluses", the number of times the y's increase as the x's increase, or the number of $y_i < y_j$ for all $i < j$,

M = "number of minuses," the number of times the y's decrease as the x's increase, or the number of $y_i > y_j$ for $i < j$.

for all $i = 1,....(n - 1)$ and $j = (i+1),.....n$.

Note that there are $n(n-1)/2$ possible comparisons to be made among the n data pairs. If all y values increased along with the x values, $S = n(n-1)/2$. In this situation, the correlation coefficient τ should equal +1. When all y values decrease with increasing x, $S = -n(n-1)/2$ and τ should equal −1. Therefore dividing S by $n(n-1)/2$ will give a value always falling between −1 and +1. This then is the definition of τ, measuring the strength of the monotonic association between two variables:

Kendall's tau correlation coefficient

$$\boxed{\tau = \frac{S}{n(n-1)/2}} \qquad [8.2]$$

To test for significance of τ, S is compared to what would be expected when the null hypothesis is true. If it is further from 0 than expected, H_0 is rejected. For $n \leq 10$ an exact test should be computed. The table of exact critical values is found in table B8 of the Appendix.

8.2.2 Large Sample Approximation

For $n > 10$ the test statistic can be modified to be closely approximated by a normal distribution. This large sample approximation Z_S is the same form of approximation as used in Chapter 5 for the rank-sum test, where now

$$d = 2 \quad \text{(S can vary only in jumps of 2)},$$
$$\mu_S = 0, \quad \text{and}$$
$$\sigma_S = \sqrt{(n/18)\bullet(n-1)\bullet(2n+5)}\ .$$

$$Z_S = \begin{cases} \dfrac{S-1}{\sigma_S} & if\ S>0 \\ 0 & if\ S=0 \\ \dfrac{S+1}{\sigma_S} & if\ S<0 \end{cases} \qquad [8.3]$$

The null hypothesis is rejected at significance level α if $|Z_S| > Z_{crit}$ where Z_{crit} is the value of the standard normal distribution with a probability of exceedance of $\alpha/2$. In the case where some of the x and/or y values are tied the formula for σ_S must be modified, as discussed in the next section.

<u>Example 1:</u> 10 pairs of x and y are given below, ordered by increasing x:

y	1.22	2.20	4.80	1.28	1.97	1.46	2.64	2.34	4.84	2.96
x	2	24	99	197	377	544	632	3452	6587	53,170

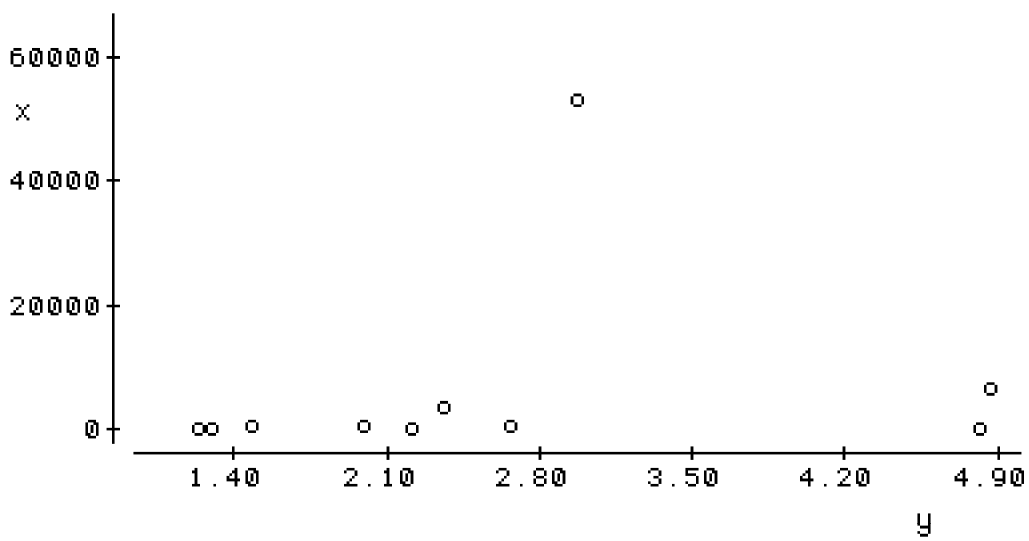


Figure 8.4 Example 1 data showing one outlier present.

To compute S, first compare y_1= 1.22 with all subsequent y's (y_j , j>1).

2.20 > 1.22, so score a +
4.80 > 1.22, score a +
1.28 > 1.22, score a +
1.97 > 1.22, score a + etc.

All subsequent y's are larger, so there are 9 +'s for i=1.

Move on to i=2, and compare y_2 =2.20 to all subsequent y's.

4.80 > 2.20, so score a +
1.28 < 2.20, score a −
1.97 < 2.20, score a −
1.46 < 2.20, score a − etc.

There are 5 +'s and 3 −'s for i=2. Continue in this way, until the final comparison of y_{n-1} = 4.84 to y_n. It is convenient to write all +'s and −'s below their respective y_i, as below:

y_i	1.22	2.20	4.80	1.28	1.97	1.46	2.64	2.34	4.84	2.96
	+	+	−	+	−	+	−	+	−	
	+	−	−	+	+	+	+	+		
	+	−	−	+	+	+	+			
	+	−	−	+	+	+				
	+	+	−	+	+					
	+	+	+	+						
	+	+	−							
	+	+								
	+									

In total there are 33 +'s (P = 33) and 12 −'s (M = 12). Therefore S = 33 − 12 = 21.
There are 10•9/2 = 45 possible comparisons, so τ = 21/45 = 0.47.
Turning to table B8, for n=10 and S=21, the exact p-value is 2•0.036 = 0.072.

The large sample approximation is

$$Z_S = (21-1) / \sqrt{(10/18)\bullet(10-1)\bullet(20+5)}$$
$$= 20/(11.18) = 1.79.$$

From a table of the normal distribution, the 1-sided quantile for 1.79 = 0.963
so that $p \cong 2 \bullet (1-.963) = 0.074$

8.2.3 Correction for Ties

To compute τ when ties are present, tied values of **either x or y** produce a 0 rather than + or − . Ties do not contribute to either P or M. S and τ are computed exactly as before. An adjustment is required for the large sample approximation Z_S , however, by correcting the σ_S formula.

In order to compute σ_S in the presence of ties, both the number of ties and the number of values involved in each tie must be counted. Consider for example a water quality data set (in units of μg/L) of 17 values (n=17) shown here in ascending order.

<1, <1, <1, <1, <1, 2, 2, 2, 3, 5, 5, 7, 9, 10, 10, 14, 18.

There are a total of 4 tied groups in the data set. The largest tied group in the data set is of 5 values (tied at <1 μg/L), there are no tied groups of 4, there is 1 tied group of 3 (at 2 μg/L), and there are 2 tied groups of 2 (at 5 and 10 μg/L). For completeness note that there are 5 "ties" of extent 1 (untied values at 3, 7, 9, 14, and 18 μg/L). These appropriately never add to the

correction because (i–1) always equals zero. Kendall (1975) defined the variable t_i as the number of ties of extent i. For this data set $t_5 = 1$ (1 tie of extent 5), $t_4 = 0$ (no ties of extent 4), $t_3 = 1$ (1 tie of extent 3), $t_2 = 2$ (2 ties of extent 2) and $t_1 = 5$ (5 "ties" of extent 1). For i>5, $t_i = 0$. Kendall's correction to σ_S in the presence of ties is:

$$\sigma_S = \sqrt{\frac{[n\,(n-1)\,(2n+5) - \sum_{i=1}^{n} t_i\,(i)\,(i-1)\,(2i+5)\,]}{18}} \qquad [8.4]$$

So for the example water quality data:

$$\sigma_S = \sqrt{[17 \bullet 16 \bullet 39 - 5 \bullet 1 \bullet 0 \bullet 7 - 2 \bullet 2 \bullet 1 \bullet 9 - 1 \bullet 3 \bullet 2 \bullet 11 - 1 \bullet 5 \bullet 4 \bullet 15] \,/\, 18}$$

or $\sigma_S = \sqrt{567} = 23.81$. Notice that if the data set could have been measured with sufficient precision (including a lower detection limit) so that no ties existed, then $\sigma_S = \sqrt{589.333} = 24.28$. Thus the ties here represent a rather small loss of information.

<u>Example 2:</u>

The example 1 data are modified to include ties, as follows:

y	1.22	2.20	4.80	1.28	1.97	1.97	2.64	2.34	4.84	2.96
x	2	24	99	99	377	544	632	3452	6587	53,170

Using a 0 to denote a tie, the comparisons used to compute P, M, and S are:

+	+	0_x	+	0_y	+	–	+	–
+	–	–	+	+	+	+	+	
+	–	–	+	+	+	+		
+	–	–	+	+	+			
+	+	–	+	+				
+	+	+	+					
+	+	–						
+	+							
+								

0_x : tie in x
0_y : tie in y

In total there are 33 +'s (P=33) and 10 –'s (M=10). Therefore S = 33–10 = 23, and $\tau = 23/45 = 0.51$. The exact two-sided p-value from table B8 is $2 \bullet 0.023 = 0.046$. For the large sample approximation, there are 2 ties of extent 2, so that

$$\sigma_S = \sqrt{[10 \bullet 9 \bullet 25 - 2 \bullet 2 \bullet 1 \bullet 9] \,/\, 18} = \sqrt{123} = 11.09$$

whereas without the tie σ_S was 11.18. Computing Zs,

$$Z_S = (23-1) \,/\, \sqrt{123}$$
$$= 22/(11.09) = 1.98.$$

From a table of the normal distribution, the 1-sided quantile for 1.98 = 0.976 so that $p \cong 2 \bullet (1-.976) = 0.048$.

8.3 Spearman's Rho

Spearman's rho is an alternative rank correlation coefficient to Kendall's tau. Kendall's tau is related to the sign test -- all positive differences between data pairs are assigned a +1 without regard to the magnitude of those differences. With Spearman's rho, differences between data values ranked further apart are given more weight, similar to the signed-rank test. Rho is perhaps easiest to understand as the linear correlation coefficient computed on the ranks of the data. Thus rho can be computed as a rank transform method. Rho and tau use different scales to measure the same correlation, much like Centigrade and Fahrenheit measures of temperature. Though tau is generally lower than rho in magnitude, their p-values for significance should be quite similar when computed on the same data.

To compute rho, the data for the two variables are ranked independently among themselves. For the ranks of x (Rx_i) and ranks of y (Ry_i), rho can be computed from the equation:

$$\text{rho} = \frac{\sum_{i=1}^{n} \left(Rx_i Ry_i\right) - n\left(\frac{n+1}{2}\right)^2}{n(n^2-1)/12} \qquad [8.5]$$

where (n+1)/2 is the mean rank of both x and y. Ties in x or y are assigned average ranks. This equation can be derived from substituting Rx_i and Ry_i for x_i and y_i in equation 8.6 for Pearson's r, and simplifying. If there is a positive correlation, the higher ranks of x will be paired with the higher ranks of y, and their product will be large. For a negative correlation the higher ranks of x will be related to lower ranks of y, and their product will be small. When there is no correlation, there will be nothing other than a random pattern in the association between x and y ranks, and their product will be similar to the product of their average rank, the second term in the numerator of equation 8.5. Thus rho will be close to zero.

Bhattacharyya and Johnson (1977) present the exact and large sample approximation versions of the hypothesis test for Spearman's rho. However, it is easiest to rank the two variables and compute the hypothesis test for Pearson's r -- the rank transform method. It is important to note that the large sample and rank approximations for rho do not fit the distribution of the test statistic well for small sample sizes (n<20), in contrast to Kendall's tau. This is one reason tau is often preferred over rho.

<u>Example 1, continued</u>

For the example 1 data, the data ranks are

Ry	1	5	9	2	4	3	7	6	10	8
Rx	1	2	3	4	5	6	7	8	9	10

Solving for rho, multiplying the ranks above gives,

$(Rx_i \bullet Ry_i)$ | 1 10 27 8 20 18 49 48 90 80, $\Sigma = 351$

$$\text{Rho} = \frac{351 - 10\,(5.5)^2}{1099/12} = \frac{48.5}{82.5}$$

= 0.588, exact p-value = 0.04 from table 13 of Bhattacharyya and Johnson (1977).

The approximate significance test for Pearson's r on the data ranks (as described in the next section) has a p-value = 0.074, not too close to the exact value. Whenever using Spearman's rho for sample sizes less than 20, exact p-values should be used.

8.4 Pearson's r

The most commonly-used measure of correlation is Pearson's r. It is also called the linear correlation coefficient because r measures the **linear** association between two variables. If the data lie exactly along a straight line with positive slope, then r = 1. This assumption of linearity makes inspection of a plot even more important for r than for rho or tau because a non-significant value of r may be due to curvature or outliers as well as to independence. As in figure 8.1, x and y may be strongly related in a nonlinear fashion, while the resulting r may be small and insignificantly different from zero.

Pearson's r is not as resistant to outliers as was tau and rho because it is computed using non-resistant measures -- means and standard deviations. It assumes that the data follow a bivariate normal distribution. With this distribution, not only do the individual variables x and y follow a normal distribution, but their joint variation also follows a specified pattern. This assumption rules out the use of r when the data have increasing variance, as in figure 8.1. Skewed variables often demonstrate outliers and increasing variance. Thus r is often not useful for describing the correlation between untransformed hydrologic variables.

Pearson's r is invariant to scale changes, as in converting streamflows in cubic feet per second into cubic meters per second, etc. This dimensionless property is obtained by standardizing, dividing the distance from the mean by the sample standard deviation, as shown in the formula for r, below.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \qquad [8.6]$$

The significance of r can be tested by determining whether r differs from zero. The test statistic t_r is computed by equation 8.7, and compared to a table of the t distribution with n−2 degrees of freedom.

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad [8.7]$$

Example 1, continued

For the example 1 data, means and standard deviations are:

	x	y
mean	6508.6	2.57
s	16531.6	1.31

Then $$r = \frac{1}{9}\sum_{i=1}^{9}\left(\frac{x_i - 6508.6}{16531.6}\right)\left(\frac{y_i - 2.57}{1.31}\right) = 0.174$$

To test for whether r is significantly different from zero, and therefore y is linearly dependent on x,

$$t_r = \frac{0.174\sqrt{8}}{\sqrt{1-(0.174)^2}} = 0.508,$$

with a p-value of 0.63 from a table of the t-distribution. Therefore H_0: r=0 is not rejected, and y is not linearly dependent (or related) to x as measured by r. This differs from the results using rho and tau, whose p-values of 0.04 and 0.07 respectively did indicate an association between y and x. Figure 8.4 provides an intuitive explanation of why r differs from rho and tau -- r is strongly affected by the one outlying observation, even though the overall trend is a linear one.

Exercises

8.1 Are uranium concentrations correlated with total dissolved solids in the following groundwater samples? If so, describe the strength of the relationship.

Uranium conc. in ppb	TDS, in mg/L	Uranium conc. in ppb	TDS, in mg/L
682.65	0.9315	1240.81	6.8559
819.12	1.9380	538.35	0.4806
303.76	0.2919	607.75	1.1452
1151.40	11.9042	705.89	6.0876
582.42	1.5674	1290.57	10.8823
1043.39	2.0623	526.09	0.1473
634.84	3.8858	784.68	2.6741
1087.25	0.9772	953.14	3.0918
1123.51	1.9354	1149.31	0.7592
688.09	0.4367	1074.22	3.7101
1174.54	10.1142	1116.59	7.2446
599.50	0.7551		

8.2 Compute the other two correlation coefficients not chosen in Exercise 8.1. Are all coefficients equally appropriate?

8.3 For the data on Corbicula densities in the Tennesse River found in Appendix C8, compute Kendall's tau for all pairs of data in the same strata and season, but one year apart. Is this correlation significant? How should this result be interpreted?

Chapter 9

Simple Linear Regression

The relationship between two continuous variables, sediment concentration and stream discharge, is to be investigated. Of interest is the quantification of this relation into a model form for use as a predictive tool during days in which discharge was measured but sediment concentration was not. Some measure of the significance of the relationship is desired so that the analyst can be assured that it is in fact composed of more than just background noise. A measure of the quality of the fit is also desired.

Sediment concentrations in an urban river are investigated to determine if installation of detention ponds throughout the city have decreased instream concentrations. Linear regression is first performed between sediment concentration and river discharge to remove the variation in concentrations which are due to flow variations. After subtracting this linear relation from the data, the residual variation before versus after the installation of ponds can be compared to determine their effect.

Regression of sediment concentration versus stream discharge is performed to obtain the slope coefficient for the relationship. This coefficient is tested to see if it is significantly different than a value obtained 5 years before using a rainfall-runoff model of the basin.

The above examples all perform a linear regression between the same two variables, sediment concentration and water discharge, but for three different objectives. Regression is commonly used for at least these three objectives. This chapter will present the assumptions, computation and applications of linear regression, as well as its limitations and common misapplications by the water resources community.

Ordinary Least Squares (OLS), commonly referred to as linear regression, is a very important tool for the statistical analysis of water resources data. It is used to describe the covariation between some variable of interest and one or more other variables. Regression is performed to

1) learn something about the relationship between the two variables, or
2) remove a portion of the variation in one variable (a portion that is not of interest) in order to gain a better understanding of some other, more interesting, portion of the variation, or
3) estimate or predict values of one variable based on knowledge of another variable, for which more data are available.

This chapter deals with the relationship between one continuous variable of interest, called the **response variable**, and one other variable -- the **explanatory variable**. The name "simple linear regression" is applied because one explanatory variable is the simplest case of regression models. The case of multiple explanatory variables is dealt with in Chapter 11 -- multiple regression.

9.1 The Linear Regression Model

The model for simple linear regression is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i=1,2,....,n$$

where

y_i	is the ith observation of the response (or dependent) variable
x_i	is the ith observation of the explanatory (or independent) variable
β_0	is the intercept
β_1	is the slope
ε_i	is the random error or residual for the ith observation, and
n	is the sample size.

The error around the linear model ε_i is a random variable. That is, its magnitude is not controlled by the analyst, but arises from the natural variability inherent in the system. ε_i has a mean of zero, and a constant variance σ^2 which does not depend on x. Due to the latter, ε_i is independent of x_i.

Regression is performed by estimating the unknown true intercept and slope β_0 and β_1 with b_0 and b_1, estimates derived from the data. As an example, in figure 9.1 the true linear relationship between an explanatory variable x and the response variable y is represented by a solid line. Around the line are 10 observed data points which result from that relationship plus the random error ε_i inherent in the natural system and the process of measurement. In practice the true line

is never known -- instead the analyst measures the 10 data points and estimates a linear relationship from those points. The OLS estimate developed from the 10 measurements is shown as the dashed line in figure 9.2.

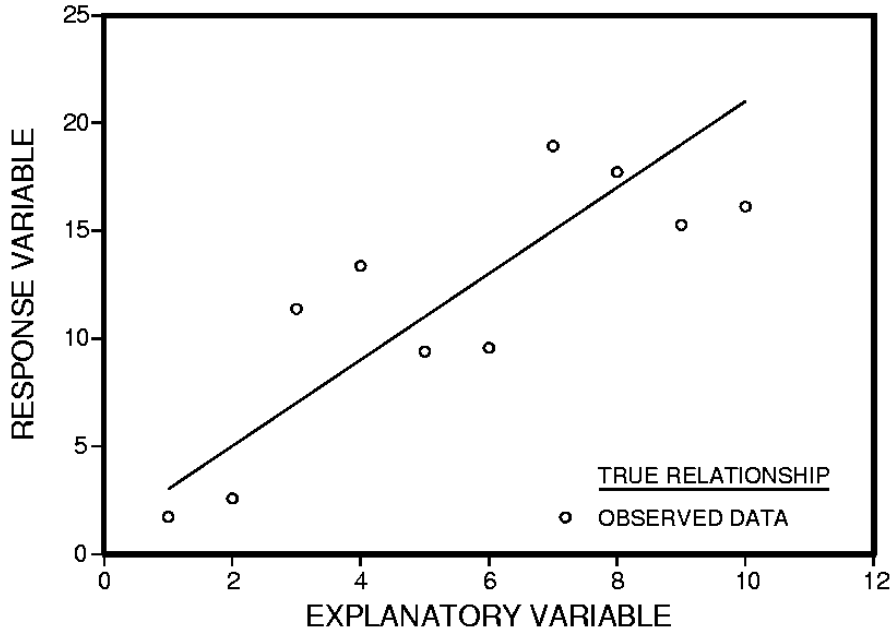


Figure 9.1 True linear relation between x and y, and 10 resultant measurements.

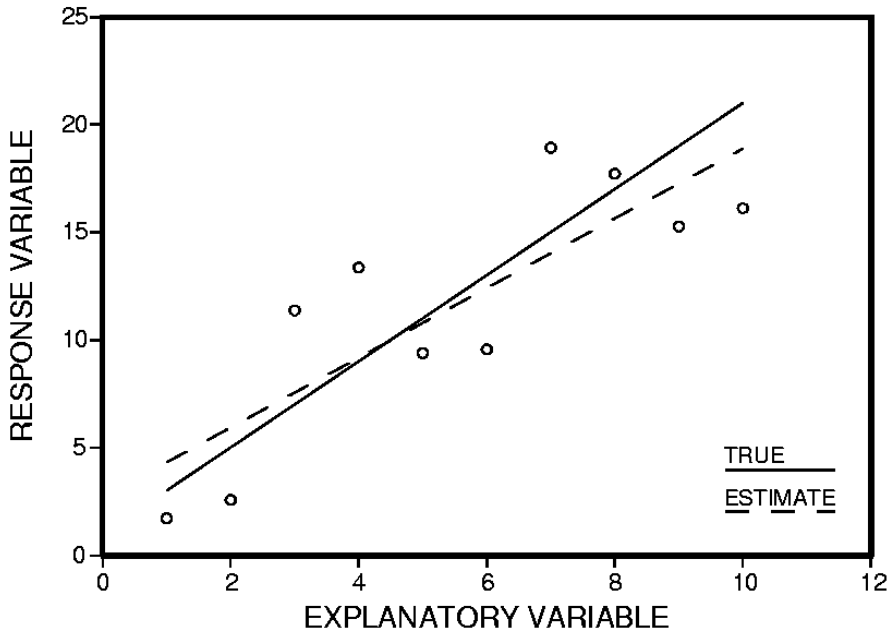


Figure 9.2 True and estimated linear relation between x and y.

If 10 new data points resulting from the same true (solid line) relationship are measured and their OLS line computed, slightly different estimates of b_0 and b_1 result. If the process is repeated several times, the results will look like figure 9.3. Some of the line estimates will fall closer to the true linear relationship than others. Therefore a regression line should always be considered as a sample estimate of the true, unknown linear relationship.

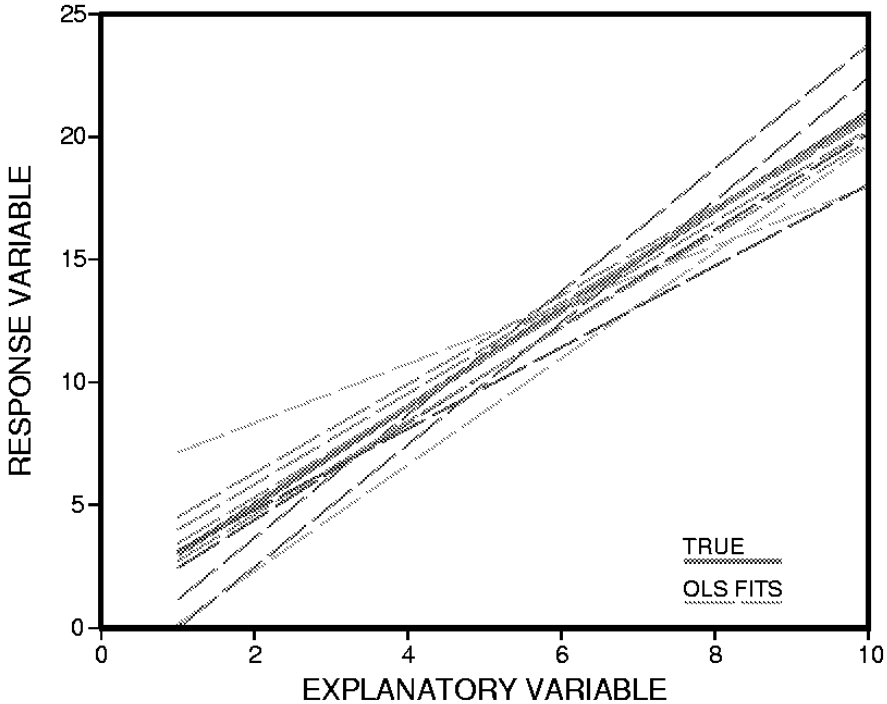


Figure 9.3 True and several estimated linear relations between x and y.

Another way of describing the linear regression model is as an estimate of the mean of y, given some particular value for x. This is called a conditional distribution. If x takes on the value x_0, then y has a conditional mean of $\beta_0 + \beta_1 x_0$ and conditional variance σ^2. The mean is "conditioned", or depends on, that particular value of x. It is the value expected for y given that x equals x_0. Therefore:

the "expected value" of y given x_0	$E[y \mid x_0]$	$= \beta_0 + \beta_1 x_0$
the variance of y given x_0	$Var[y \mid x_0]$	$= \sigma^2$

9.1.1 Assumptions of Linear Regression

There are five assumptions associated with linear regression. These are listed in table 9.1. The necessity of satisfying them is determined by the purpose to be made of the regression equation. Table 9.1 indicates for which purposes each is needed.

	Purpose			
Assumption	Predict y given x	Predict y and a variance for the prediction	Obtain best linear unbiased estimator of y	Test hypotheses, estimate confidence or prediction intervals
(1) Model form is correct: y is linearly related to x	+	+	+	+
(2) Data used to fit the model are representative of data of interest.	+	+	+	+
(3) Variance of the residuals is constant (is homoscedastic). It does not depend on x or on anything else (e.g. time).		+	+	+
(4) The residuals are independent.			+	+
(5) The residuals are normally distributed.				+

Table 9.1 Assumptions necessary for the purposes to which OLS is put.
+: the assumption is required for that purpose.

The assumption of a normal distribution is involved only when testing hypotheses, requiring the residuals from the regression equation to be normally distributed. In this sense OLS is a parametric procedure. No assumptions are made concerning the distributions of either the explanatory or response variables. The most important hypothesis test in regression is whether the slope coefficient is significantly different from zero. Normality of residuals is required for this test, and should be checked by a boxplot or probability plot. The regression line, as a conditional mean, is sensitive to the presence of outliers in much the same way as a sample mean is sensitive to outliers.

9.2 Computations

Linear regression estimation is nothing more than a minimization problem. It can be stated as follows: find two numbers b_0 and b_1 such that

$\sum_{i=1}^{n} (y_i - y_i)^2$ is minimized, where $\hat{y}_1$ is the OLS estimate of y:

$$\hat{y}_i = b_0 + b_1 x_i.$$

This can be solved for b_0 and b_1 using calculus. The solution is referred to as the normal equations. From these come an extensive list of expressions used in regression:

Formula	Name
$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$	mean x
$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$	mean y
$SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$	sums of squares y = Total SS
$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$	sums of squares x
$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i y_i) - n\,\bar{x}\,\bar{y}$	sums of x y cross products
$b_1 = S_{xy} / SS_x$	the estimate of β_1 (slope)
$b_0 = \bar{y} - b_1\,\bar{x}$	the estimate of β_0 (intercept)
$\hat{y}_i = b_0 + b_1 x_i$	the estimate of y given x_i

Formula	Name
$e_i = y_i - \hat{y}_i$	the estimated residual for obs. i
$SSE = \sum_{i=1}^{n} e_i^2$	error sum of squares
$s^2 = (SS_y - b_1 S_{xy}) / (n-2)$ $= \sum_{i=1}^{n} e_i^2 / (n-2)$	The estimate of σ^2, also called mean square error (MSE).
$s = \sqrt{s^2}$	standard error of the regression or standard deviation of residuals
$SE(\beta_1) = s / \sqrt{SS_x}$	standard error of β_1
$SE(\beta_0) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}$	standard error of β_0
$r = S_{xy} / \sqrt{SS_x SS_y}$ $= b_1 \sqrt{SS_x / SS_y}$	the correlation coefficient
$R^2 = [SS_y - s^2 (n-2)] / SS_y$ $= 1 - (SSE / SS_y)$ $= r^2$	coefficient of determination, or fraction of the variance explained by regression

9.2.1 Properties of Least Squares Solutions

1) If assumptions 1 through 4 are all met, then the estimators b_0 and b_1 are the minimum variance unbiased estimators of β_0 and β_1.
2) The mean of the residuals (e_i's) is exactly zero.
3) The mean of the predictions ($\hat{y}_i$'s) equals the mean of the observed responses (y_i's).
4) The regression line passes through the centroid of the data ($\bar{x}$, $\bar{y}$).
5) The variance of the predictions ($\hat{y}_i$'s) is less than the variance of the observed responses (y_i's) unless $R^2 = 1.0$.

9.3 Building a Good Regression Model

A common first step in performing regression is to plug the data into a statistics software package and evaluate the results using R^2. Values of R^2 close to 1 are often incorrectly deemed an indicator of a good model. This is a dangerous, blind reliance on the computer software. An R^2 near 1 can result from a poor regression model; lower R^2 models may often be preferable. Instead of the above, performing the following steps in order will generally lead to a good regression model.

The following sections will use the total dissolved solids (TDS) concentrations from the Cuyahoga River at Independence, Ohio, 1974-1985 as an example data set. The data are found in Appendix C9. These concentrations will be related to stream discharge (Q).

1) First step -- PLOT THE DATA!
   Plot y versus x and check for two things
   1a) does the relationship look non-linear?
   1b) does the variability of y look markedly different for different levels of x?

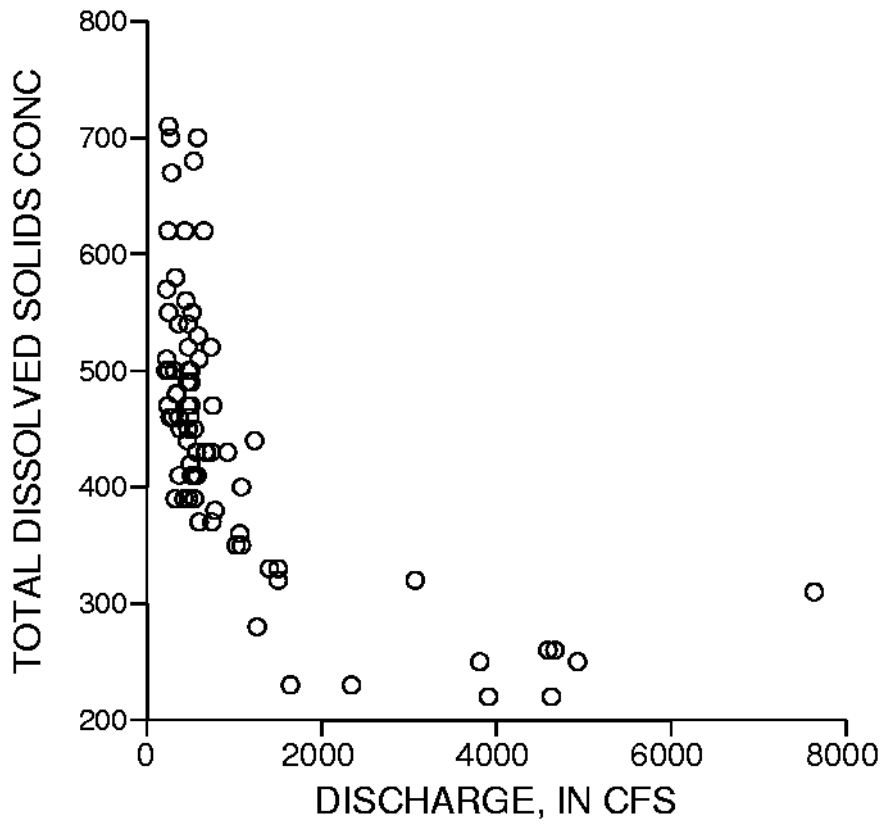


Figure 9.4 Scatterplot of the Cuyahoga R. TDS data

If the problem is curvature only (1a), then try to identify a new x which is a better linear predictor (a transform of the original x or another variable altogether). When possible, use the

best physically-based argument in choosing the right x. It may be appropriate to resort to empirically selecting the x which works best (highest R^2) from among a set of reasonable explanatory variables.

If the problem is non-constant variance, (also called heteroscedasticity, 1b above) or both curvature and heteroscedasticity, then transforming y, or x and y, may be called for. Mosteller and Tukey (1977) provided a guide to selecting power transformations using plots of y versus x called the "bulging rule". Going "up" the ladder of powers means $\theta > 1$ (x^2, etc.) and "down" the ladder of powers means $\theta < 1$ (log x, 1/x, $\sqrt{x}$, etc.).

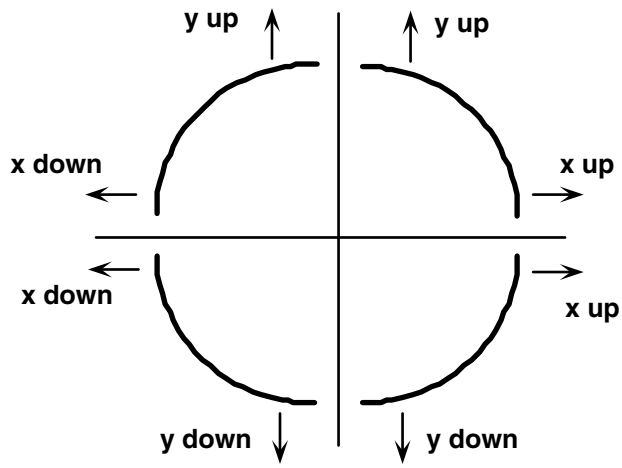


Figure 9.5 The bulging rule for transforming curvature to linearity.
(after Mosteller and Tukey, 1977).

The non-linearity of the TDS data is obvious from figure 9.4, and some type of transformation of the x variable (discharge, denoted Q) is necessary. The base 10 log of Q is chosen, as the plot has the shape of the lower left quadrant of the bulging rule, and so $\theta < 1$. Figure 9.6 presents the TDS data versus the log of Q. Linearity is achieved. There is some hint of greater variance around the line at the lower Q's, but notice that there are also far more data at lower discharges. The range of values can be expected to be greater where there is more data, so non-constant variance is not proven. Therefore this transformation appears acceptable based on the first set of plots.

2) Having selected an appropriate x and y, compute the least squares regression statistics, saving the values of the residuals for further examination. In the regression results, focus on these things:

2a) The coefficients, b_0 and b_1: Are they reasonable in sign and magnitude? Do they lead to predictions of unreasonable values of y for reasonable values of x (e.g., negative flows or concentrations)?

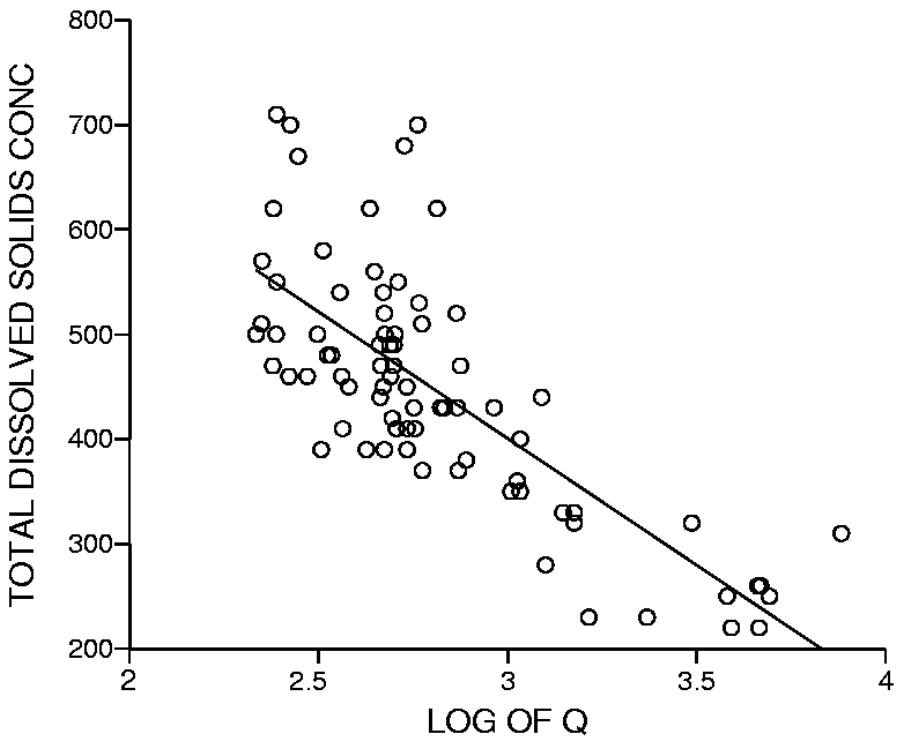


Figure 9.6 Scatterplot with regression line after transformation of x

The Cuyahoga TDS data have the following regression results:

$TDS = 1125 - 242 \log_{10} Q$

$n = 80$ $s = 75.55$ $R^2 = 0.57$ $SS_x = 10.23$

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	1125.5	66.9	16.8	0.000
Slope β_1	−241.6	23.6	−10.2	0.000

Table 9.2 Regression statistics for the Cuyahoga TDS data

It appears reasonable that TDS concentrations should be diluted with increasing stream discharge, producing a negative slope. No negative concentrations result from reasonable values for Q at this site.

2b) The R^2: Does the regression explain much variance? Is the amount of variance explained substantial enough to make it worthwhile to use the regression, given the risk that the form of the model is likely to be imperfect? There is no general rule for what is too low an R^2 for a useful regression equation.

For the Cuyahoga data, 57% of the variance of total dissolved solids is explained by the effect of log Q.

2c) Look at the t-ratio (or t-statistics) on the two coefficients. These are the test statistics needed for testing the null hypothesis that the coefficient is equal to zero. In particular, look at the t-ratio on β_1. If $|t|>2$, reject $\beta_1=0$ at $\alpha=0.05$ for reasonably large sample sizes and therefore assert there is a statistically significant linear relationship between x and y. If the t-ratio is between −2 and +2, the observed relationship is no stronger than what is likely to arise by chance alone in the absence of any real linear relationship. If this is the case one should go back to step 1 or give up on the use of regression with this data set. (The formalities of these hypothesis tests are given in a later section).
Both the intercept and slope of the TDS regression are significant at any reasonable α, as shown by the large t-statistics and small p-values of table 9.2.

3) Examine adherence to the assumptions of regression using residuals plots. Three types of residuals plots will clearly present whether or not the regression model adheres sufficiently to the assumptions to be used.

3a) Residuals versus predicted (e vs. $\hat{y}$). Look for two possible problems: curvature and heteroscedasticity. These are exactly the same problems described in step 1. However, plotting residuals enhances the opportunity to see these problems as compared to plotting the original data. The solutions to the problems are the same. Figure 9.7 presents an example of a good residuals plot, one where the residuals show no curvature or changing variance. Figure 9.8, on the other hand, is a residuals plot which shows both curvature and changing variance, producing the typical "horn" pattern which is often correctable by taking the logarithms of y.

It is possible to read too much into these plots, however. Beware of "curvature" produced by a couple of odd points or of error variance seeming to both grow and shrink one or more times over the range of $\hat{y}$. Probably neither of these can or should be fixed by transformation but may indicate the need for the robust procedures of Chapter 10.

In figure 9.9, the residuals from the Cuyahoga TDS regression are plotted versus its predicted values. There is an indication of heteroscedasticity, though again there are more data for the larger predicted values. There also appears to be a bow in the data, from + to − and back to + residuals. Perhaps a transformation of the TDS concentrations are warranted, or the incorporation of additional variables into the regression equation.

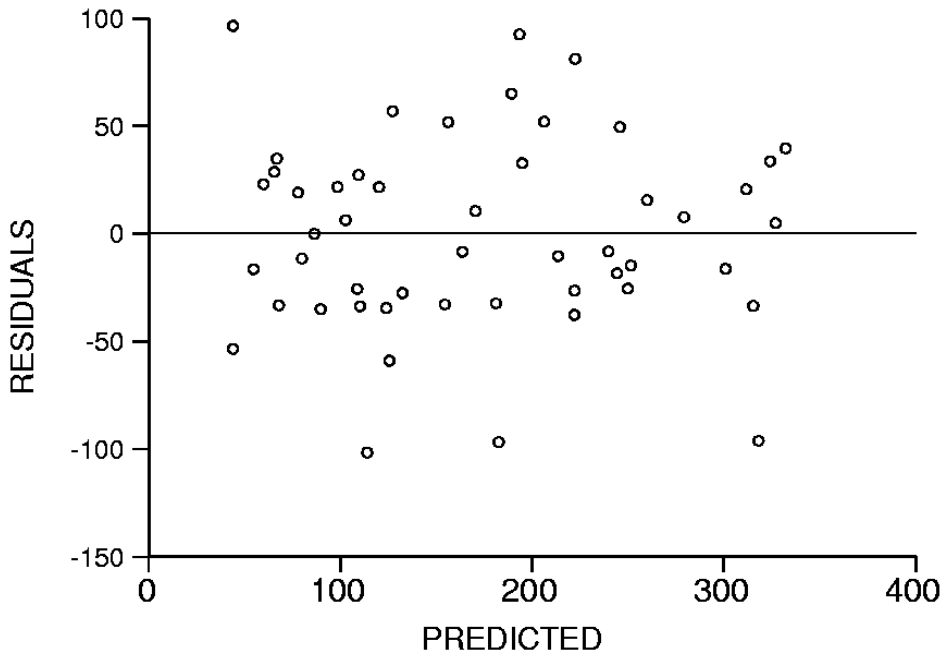


Figure 9.7 Example of a residuals plot for a good regression model

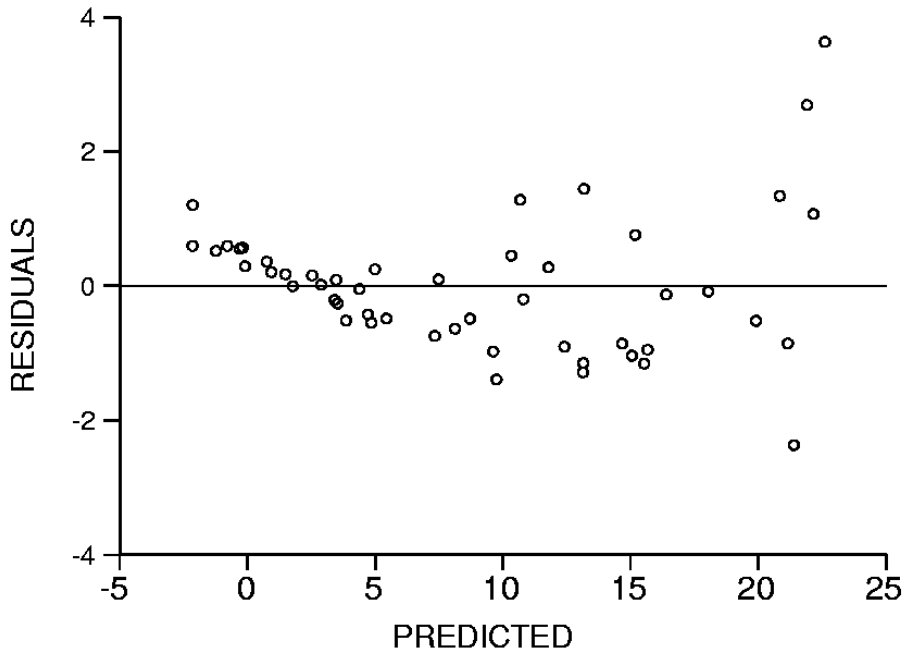


Figure 9.8 Residuals plot showing curvature and changing variance.

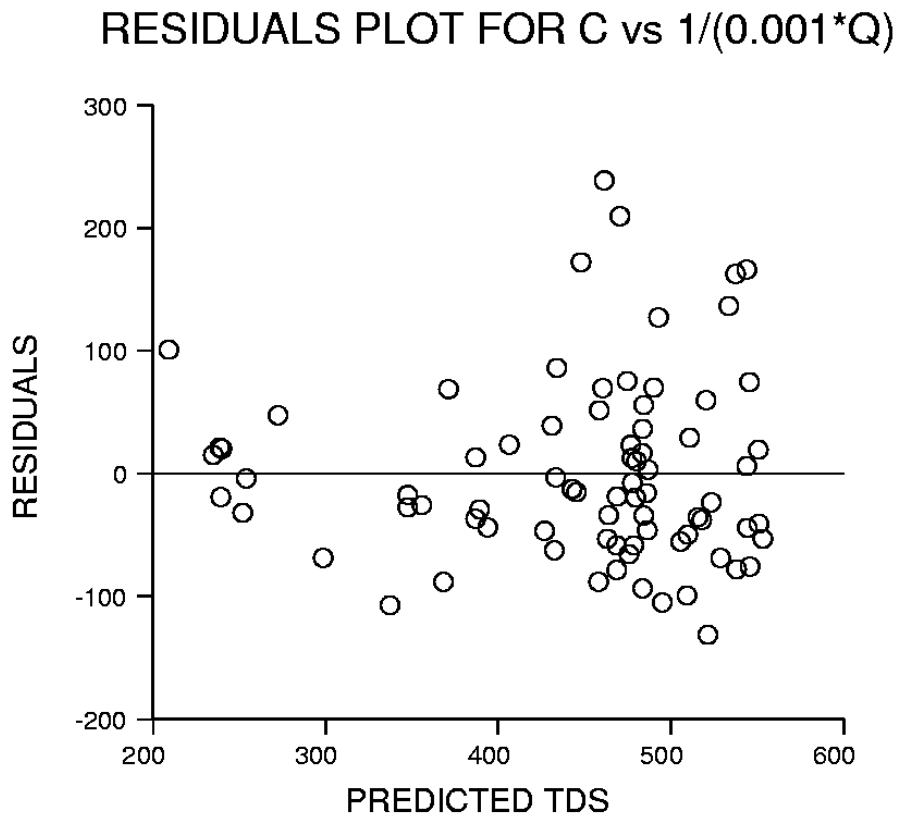


Figure 9.9 Residuals plot of the Cuyahoga data.

3b) Residuals versus time (e vs. t). If there is any time or space order to the observations (relating to time of collection, time of measurement, or map location), plot the residuals versus time or season or time of day, or versus the appropriate 1- or 2-dimensional space coordinate to see if there is a pattern in the residuals. A good residuals pattern, one with no relation between residuals and time, will look similar to figure 9.7 -- random noise. If on the other hand structure in the pattern over time is evident, seasonality, long-term trend, or correlation in the residuals may be the cause. Trend or seasonality suggest adding a new term to the regression equation (see Chapter 12). Correlation between residuals over time or space require one of the remedies listed in section 9.5.4.

Correlation between residuals over time or space may not be evident from the e_i versus $\hat{y}$ residuals plot (figure 9.10a), but will stand out on a plot of e_i versus time (9.10b). The nonrandomness is evident in that positive residuals clump together, as do negative -- a positive correlation. Plotting the ith versus the (i–1)th residual shows this pattern more strongly (9.10c). If time or space are measured as categorical variables (month, etc.), plot boxplots of residuals by category and look for patterns of regularity. Where no differences occur between boxes, the time or space variable has no effect on the response variable.

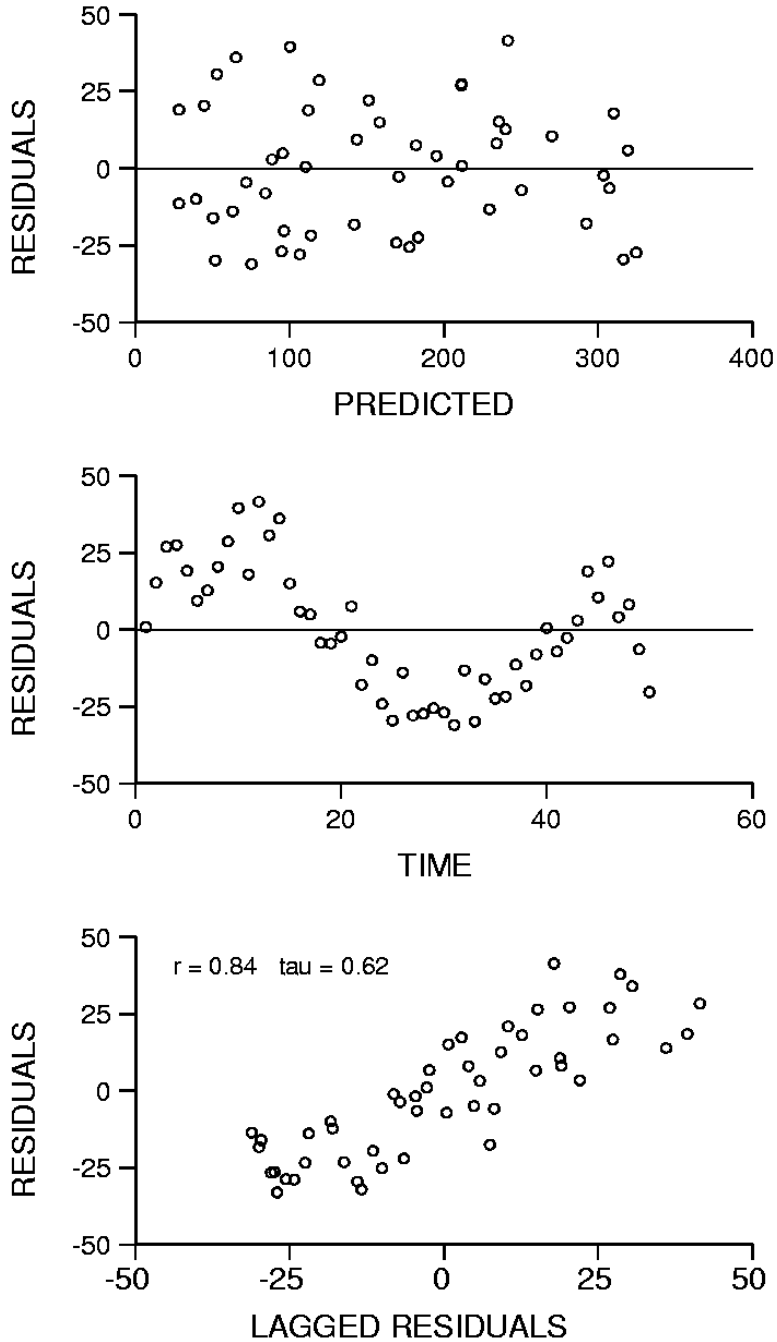


Figure 9.10 a) Residual e_i vs $\hat{y}$ plot shows no hint of correlation over time
b) Time series of residuals shows e_i related to time
c) Correlation of e_i vs. $e_{(i-1)}$

In figure 9.11, boxplots of TDS residuals by month show a definite seasonality, with generally high residuals occurring in the winter months, low residuals in the summer, and unusually high values in September. Thus the regression equation will underpredict concentrations in the

winter and overpredict in the summer. This pattern may be due to washoff of road de-icing salts in the winter. The unknown cause of the September anomalies should be investigated further. To better mimic the seasonal variation, other explanatory variables must be added. This will be discussed in Chapter 12.

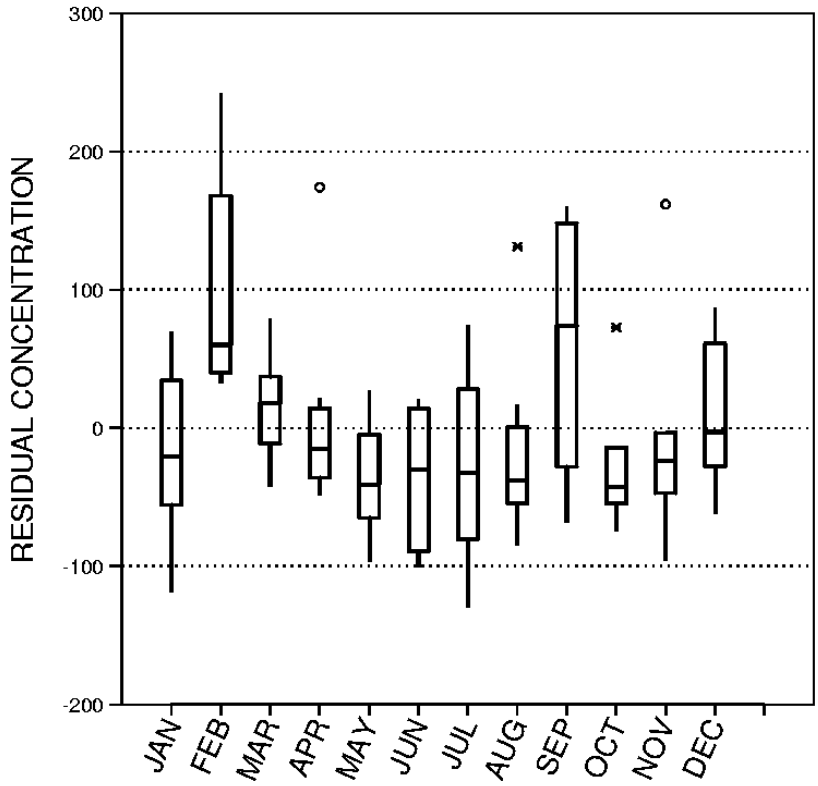


Figure 9.11 Residual of TDS concentrations by month. Note the seasonality.

3c) Normality of residuals. Examine the distribution of residuals using a boxplot, stem and leaf, histogram, or normal probability plot. If they depart very much from a normal distribution, then the various confidence intervals, prediction intervals, and tests described below will be inappropriate. Specifically,
(i) hypothesis tests will have low power (slopes or explanatory variables will falsely be declared insignificant), and
(ii) confidence or prediction intervals will be too wide, as well as giving a false impression of symmetry.

A boxplot of residuals from the TDS-logQ regression shown in figure 9.12 is mildly right-skewed, with several outliers present. A probability plot of the residuals (figure 9.13) shows a slight departure from normality. If these were the only problems, transformation of the y variable might not be warranted. But combined with the

problems already noted above of curvature and heteroscedasticity, further work is required.

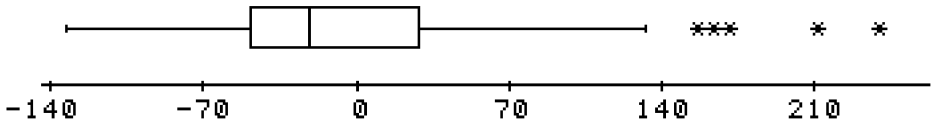


Figure 9.12 Boxplot of the TDS regression residuals

For further attempts to find an appropriate transformation of the Cuyahoga data, see problem 9.1 at the end of this chapter.

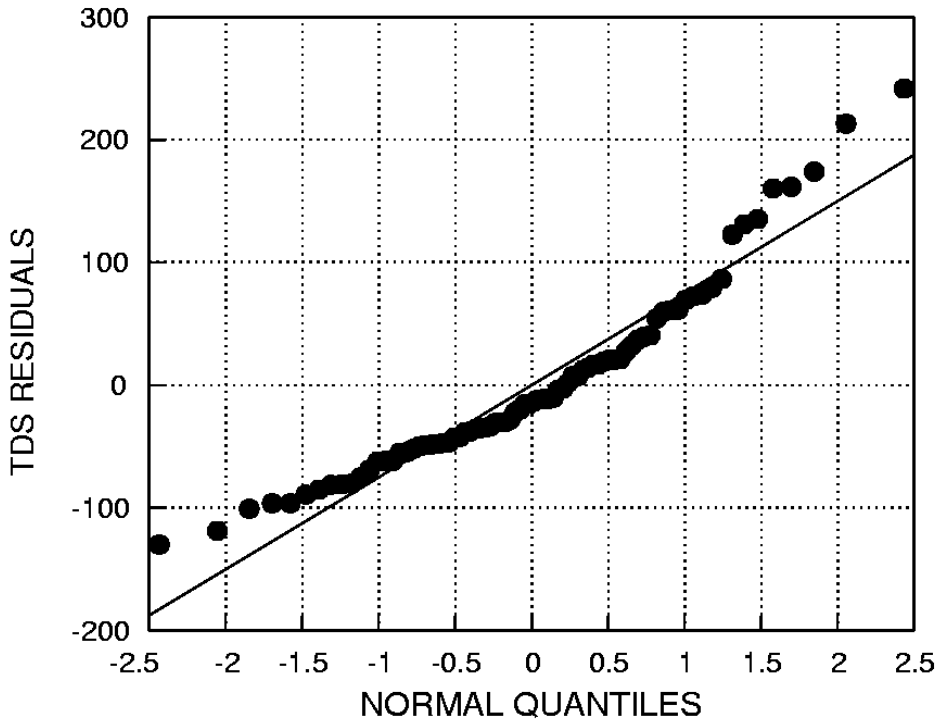


Figure 9.13 Probability plot of the TDS regression residuals

3d) Residuals versus other explanatory variables. To determine whether other explanatory variables should be included into a multiple regression model, boxplots of residuals by categorical explanatory variables or scatterplots versus continuous variables should be plotted. If something other than a random pattern occurs, that variable or one like it may be appropriate for adding to the regression equation. Figure 9.14 for example might result from plotting residuals from a regression of radon concentrations in water versus uranium content of rocks. using different symbols for wells and springs. The residuals for wells tend to be larger than those for springs, as also shown by the boxplots at the side. Incorporating an additional explanatory variable for "water source" into the

regression equation using the techniques of Chapter 11 explains more of the noise in the data, improving the model.

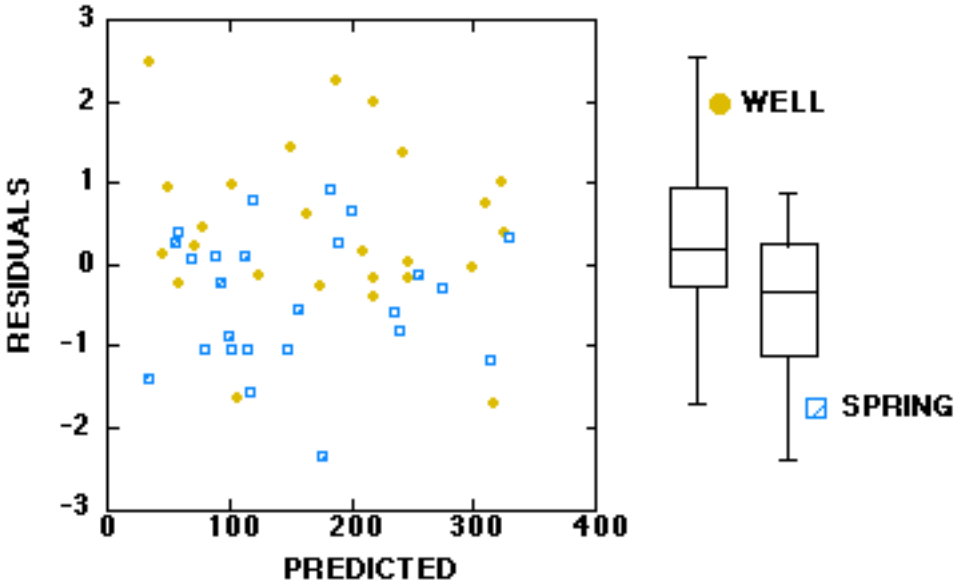


Figure 9.14 Residuals plotted by an additional explanatory variable.

4) Use the regression diagnostics of section 9.5 to ensure that one or two observations are not strongly influencing the values of the coefficients, and to determine the quality of predicted values. These diagnostics duplicate much of what can be seen with plots for a single explanatory variable, but become much more important when performing multiple regression.

9.4 Hypothesis Testing in Regression

9.4.1 Test for Whether the Slope Differs from Zero

The hypothesis test of greatest interest in regression is the test for a significant slope (β_1). Typically, the null hypothesis is

H_0: $\beta_1 = 0$

versus the alternative hypothesis

H_1: $\beta_1 \neq 0$.

The null hypothesis also states that the value of y does not vary as a linear function of x. Thus for the case of a single explanatory variable this also tests for whether the regression model has statistical significance. A third interpretation is as a test for whether the linear correlation coefficient significantly differs from zero. The latter two interpretations are <u>not</u> applicable for

multiple explanatory variables. The test statistic computed is the t-ratio (the fitted coefficient divided by its standard error):

$$t = \frac{b_1}{s/\sqrt{SS_x}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

H_0 is rejected if $|t| > t_{crit}$, where t_{crit} is the point on the Student's t distribution with n–2 degrees of freedom, and with probability of exceedance of $\alpha/2$.

Note that when $\alpha=0.05$ and $n>30$ $t_{crit} \cong 2.0$

and when $\alpha=0.01$ and $n>30$ $t_{crit} \cong 2.6$.

For the Cuyahoga TDS example the t-statistic for β_1 was much greater than 2, and indeed was significant at the $\alpha = 0.0001$ level. Therefore a strong linear correlation exists between TDS and $\log_{10}$ of Q.

This test for nonzero slope can also be generalized to testing the null hypothesis that $\beta_1=\beta_1^*$ where β_1^* is some pre-specified value. For this test the statistic is defined as

$$t = \frac{b_1 - b_1^*}{s/\sqrt{SS_x}}$$

9.4.2 Test for Whether the Intercept Differs from Zero

Tests on the intercept b_0 can also be computed. The test for

H_0: $b_0 = 0$

is usually the one of interest. The test statistic is

$$t = \frac{b_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}}$$

H_0 is rejected if $|t| > t_{crit}$ where t_{crit} is defined as in the previous test. From table 9.2 the intercept for the TDS data is seen to be highly significantly different from 0.

It can be dangerous to delete the intercept term from a regression model. Even when the intercept is not significantly different from zero, there is little benefit to forcing it to equal zero, and potentially great harm in doing so. Regression statistics such as R^2 and the t-ratio for β_1 lose their usual meaning when the intercept term is dropped (set equal to zero). Recognition of a physical reason why y must be zero when x is zero is not a sufficient argument for setting b_0 = 0. Probably the only appropriate situation for fitting a no-intercept model is when all of the following conditions are met:

1) the x data cover several orders of magnitude,
2) the relationship clearly looks linear from zero to the most extreme x values,
3) the null hypothesis that $\beta_0 = 0$ is not rejected, and
4) there is some economic or scientific benefit to dropping the intercept.

9.4.3 Confidence Intervals on Parameters

Confidence intervals for the individual parameters β_0, β_1, and σ^2 indicate how well these can be estimated. The meaning of the $(1-\alpha)\bullet 100\%$ confidence interval is that, in repeated collection of new data and subsequent regressions, the frequency with which the true parameter value would fall outside the confidence interval is α. For example, $\alpha = 0.05$ confidence intervals around the estimated slopes of the regression lines in figure 9.3 would include the true slope 95% of the time.

For the slope β_1 the confidence interval (C.I.) is

$$\left(b_1 - \frac{t\,s}{\sqrt{SS_x}} \;,\; b_1 + \frac{t\,s}{\sqrt{SS_x}} \right)$$

where t is the point on the student's t-distribution having n−2 degrees of freedom with a probability of exceedance of $\alpha/2$.

For the intercept β_0 the C.I. is

$$\left(b_0 - ts\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}, b_0 + ts\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}} \right)$$

where t is defined as above.

For the variance σ^2 (also called the mean square error MSE), the C.I. is

$$\left(\frac{(n-2)s^2}{\chi^2_{1-a/2}}, \frac{(n-2)s^2}{\chi^2_{a/2}} \right)$$

where χ^2_p is the quantile of the chi-square distribution having n−2 degrees of freedom with exceedance probability of p.

As an example, the 95% confidence intervals for the Cuyahoga TDS data are:

For β_1: $\left(-241.6 - \frac{1.99\bullet 75.6}{\sqrt{10.23}} \;,\; -241.6 + \frac{1.99\bullet 75.6}{\sqrt{10.23}} \right) = (-288.6\,,\, -194.6)$

For β_0: $\left(1125.5 - 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{2.81^2}{10.23}} \;,\; 1125.5 + 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{2.81^2}{10.23}} \right)$

$= (991.8\,,\, 1258.7)$

For σ^2: $\left(\frac{(78)\ 5708}{104.3} , \frac{(78)\ 5708}{55.5} \right) = (4269 , 8022)$

9.4.4 Confidence Intervals for the Mean Response

There is also a confidence interval for the conditional mean of y given any value of x. If x_0 is a specified value of x, then the estimate of the expected value of y at x_0 is $\hat{y} = b_0 + b_1 x_0$, the value predicted from the regression equation. But there is some uncertainty to this, associated with the uncertainty for the true parameters β_0 and β_1. The $(1-\alpha)\bullet 100\%$ confidence interval for the mean y is then

$$\left(\hat{y} - ts\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}},\ \hat{y} + ts\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}} \right)$$

where t is the quantile of the students' t-distribution having n–2 degrees of freedom with probability of exceedance of $\alpha/2$. Note that the confidence interval is two-sided, requiring a t-statistic of $\alpha/2$ for either side. Also note from the formula that the farther x_0 is from $\bar{x}$ the wider the interval becomes. That is, the model is always "better" near the middle of the x values than at the extremes.

To continue the Cuyahoga TDS example, the confidence interval for the mean y is calculated for two values of x_0, 3.0 (near $\bar{x}$) and 3.8 (far from $\bar{x}$):

for $x_0 = 3.0$: $\left(399 - 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{(3.0\text{-}2.81)^2}{10.23}} , 399 + 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{(3.0\text{-}2.81)^2}{10.23}} \right)$
$= (380 , 418)$

for $x_0 = 3.8$: $\left(205.4 - 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{(3.8\text{-}2.81)^2}{10.23}} , 205.4 + 1.99\bullet 75.6\sqrt{\frac{1}{80} + \frac{(3.8\text{-}2.81)^2}{10.23}} \right)$
$= (155.9, 254.9)$

a confidence interval of width 38 at $x_0 = 3.0$, and a width of 99 at $x_0 = 3.8$.

When the confidence interval for each logQ value is connected together, the characteristic "bow" shape of regression confidence intervals can be seen (figure 9.15). Note that this shape agrees with the pattern seen in figure 9.3 for randomly generated regression lines, where the positions of the line estimates are more tightly controlled near the center than near the ends.

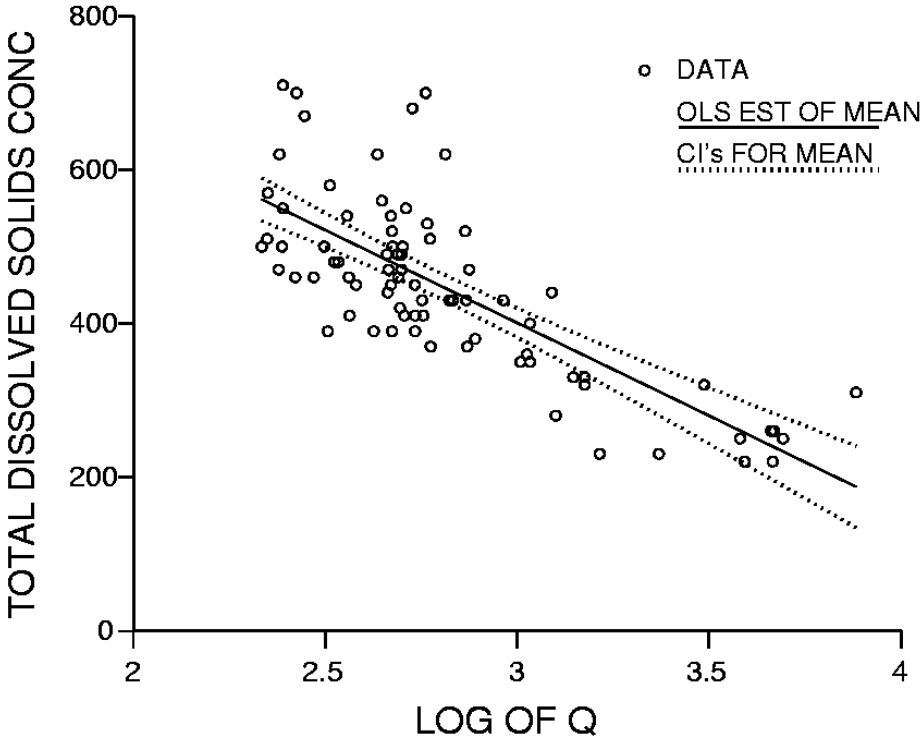


Figure 9.15 Confidence intervals for mean TDS for the Cuyahoga River data.

9.4.5 Prediction Intervals for Individual Estimates of y

The prediction interval, the confidence interval for prediction of an estimate of an individual y, is often confused with the confidence interval for the mean. This is not surprising, as the best estimate for both the mean of y given x_0 and for an individual y given x_0 are the same -- $\hat{y}$. However, their confidence intervals differ. The formulas are identical except for one very important term. The prediction interval incorporates the unexplained variability of y (σ^2) in addition to uncertainties in the parameter estimates β_1 and β_2. The $(1-\alpha)\bullet 100\%$ prediction interval for a single response is

$$\left(\hat{y} - ts\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}}, \quad \hat{y} + ts\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_x}} \right)$$

where all of the terms are as defined previously. Note that these intervals widen as x_0 departs from x, but not nearly as markedly as the confidence intervals do. In fact, a simple rough approximation to the prediction interval is just ($\hat{y}$ − ts, $\hat{y}$ + ts), two parallel straight lines. This is because the second and third terms inside the square root are negligible in comparison to the first, provided the sample size is large. These prediction intervals should contain approximately $1-\alpha\bullet(100)\%$ of the data within them, with $\alpha/2\bullet(100)\%$ of the data beyond each side of the intervals. They will do so if the residuals are approximately normal.

The prediction intervals for the Cuyahoga TDS data are plotted in figure 9.16. They are computed below for x_0 = 3.0 and 3.8.

for x_0 = 3.0:

$$\left(399 - 1.99 \bullet 75.6\sqrt{1 + \frac{1}{80} + \frac{(3.0-2.81)^2}{10.23}}\ ,\ 399 + 1.99 \bullet 75.6\sqrt{1 + \frac{1}{80} + \frac{(3.0-2.81)^2}{10.23}}\right)$$
$$= (247.4\ ,\ 550.6)$$

for x_0 = 3.8:

$$\left(205.4 - 1.99 \bullet 75.6\sqrt{1 + \frac{1}{80} + \frac{(3.8-2.81)^2}{10.23}}\ ,\ 205.4 + 1.99 \bullet 75.6\sqrt{1 + \frac{1}{80} + \frac{(3.8-2.81)^2}{10.23}}\right)$$
$$= (47.0,\ 363.8)$$

a prediction interval of width = 303 at x_0 = 3.0, and a width of 317 at x_0 = 3.8. Note that the prediction intervals are much wider than the confidence intervals, and that there is only a small difference in width between the two prediction intervals as x_0 changes. Also note from figure 9.16 that the data appear skewed, with all of the values found beyond the prediction intervals falling above the upper interval.

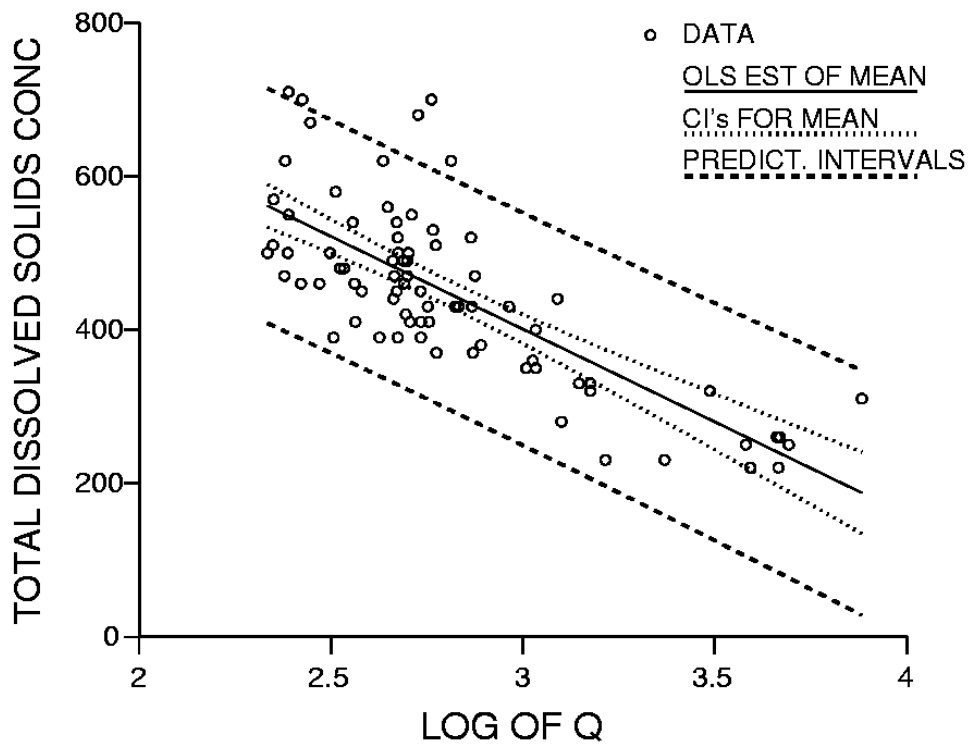


Figure 9.16 Prediction intervals for an individual TDS estimate -- Cuyahoga River.

9.4.5.1 Nonparametric prediction interval

There is also a nonparametric version of the prediction interval. This might be used when the x,y data display a linear relationship and residuals have constant variance (homoscedastic), but the distribution of the residuals appears non-normal. Typically, such departures from normality take the form of skewness or an excessive number of outside or far outside values (as seen in a boxplot). The nonparametric prediction interval is

$$(\hat{y} + e_{(L)},\ \hat{y} + e_{(U)})$$

where $e_{(L)}$ and $e_{(U)}$ are the $1-\alpha/2$ and $\alpha/2$th quantiles of the residuals.

In other words, $e_{(L)}$ is the Lth ranked residual and $e_{(U)}$ is the Uth ranked residual, where $L = (n+1)\bullet\alpha/2$ and $U = (n+1)\bullet(1-\alpha/2)$. When L and U are not integers either the integer closest to L and U can be chosen, or $e_{(L)}$ and $e_{(U)}$ can be interpolated between adjacent residuals.

For the Cuyahoga TDS data, L= 81•.025 = 2.025 and U= 81•.975 = 78.975. Either the 2nd and 79th ranked residual can be selected, or values interpolated between the 2nd and 3rd, and the 78th and 79th residual. These are then added to the regression line ($\hat{y}$). In figure 9.17 the nonparametric prediction interval is compared to the one previously developed assuming normality of residuals. Note that the nonparametric interval is asymmetric around the central regression line, reflecting the asymmetry of the data.

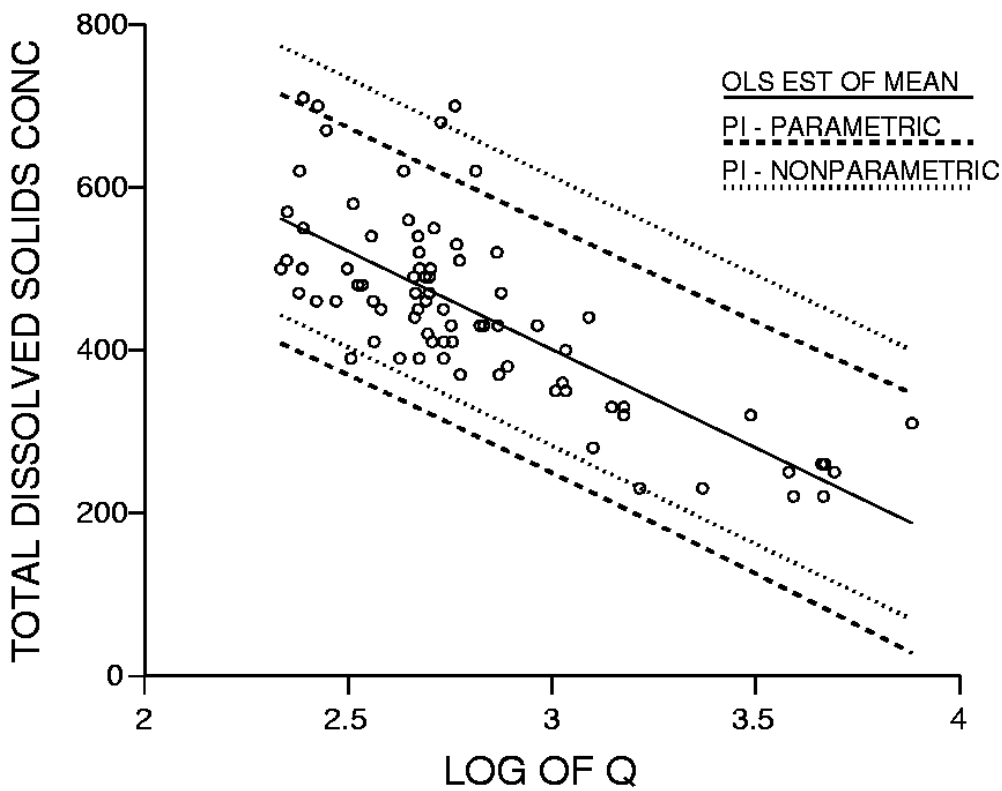


Figure 9.17 Nonparametric and parametric prediction intervals for the TDS data.

9.5 Regression Diagnostics

One common mistake in regression analysis is to base decisions about model adequacy solely on the regression summary statistics--principally R^2, s and the F- or t-test results. R^2 is a measure of the percent of the variation in the response (y) variable that is accounted for by the variation in the explanatory variables. The s (standard error of the regression or standard deviation of the residuals) is a measure of the dispersion of the data around the regression line. Most regression programs also perform an overall F-test to determine if the regression relationship is statistically significant, ie. that the apparent relationship between y and x is not likely to arise due to chance alone. Some programs also do a t-test for each explanatory variable to determine if the coefficient for that variable is significantly different from zero.

These statistics provide substantial information about regression results. An equation that accounts for a large amount of the variation in the response variable and has coefficients that are statistically significant is highly desirable. However, decisions about model adequacy cannot be made on the basis of these criteria alone. A large R^2 or significant F-statistic does not guarantee that the data have been fitted well. Figure 9.18 (Anscombe, 1973) illustrates this point.

The data in the four graphs have exactly the same summary statistics and regression line (same b_0, b_1, s, R^2). In 9.18a is a perfectly reasonable regression model, an evidently linear relationship having an even distribution of data around the least-squares line. The strong curvature in 9.18b suggests that a linear model is highly inadequate and that some transformation of x would be a better explanatory variable, or that an additional explanatory variable is required. With these improvements perhaps all of the variance could be explained. Figure 9.18c illustrates the effect of a single outlier on regression. The line mis-fits the data, and is drawn towards the outlier. Such an outlier must be recognized and carefully examined to verify its accuracy if possible. If it is impossible to demonstrate that the point is erroneous, a more robust procedure than regression should be utilized (see Chapter 10). The regression slope in 9.18d is strongly affected by a single point (the high x value), with the regression simply connecting two "points", a single point plus a small cluster of points. Such situations often produce R^2 values close to 1, yet may have little if any predictive power. Had the outlying point been in a different location, the resulting slope would be totally different. For example, the only difference between the data of figure 9.19a and 9.19b is the rightmost data point. Yet the slopes are entirely different! Regression should not be used in this case because there is no possible way to evaluate the assumptions of linearity or homoscedasticity without collecting more data in the gap between the point and cluster. In addition, the slope and R^2 are totally controlled by the position of one point, an unstable situation.

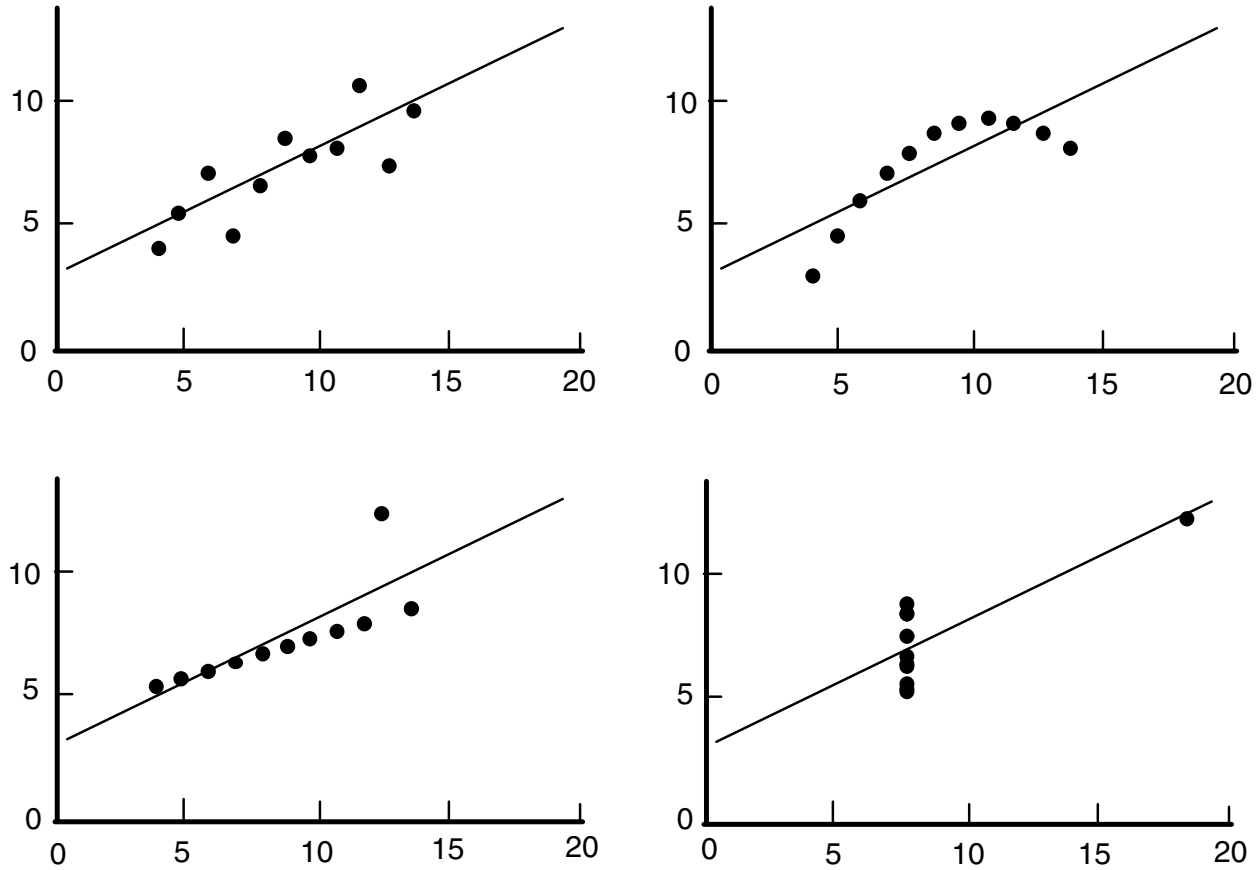


Figure 9.18 Three key pathologies in regression (after Anscombe, 1973).

© American Statistical Association. Used with permission.

The three key pathologies can be referred to by simple names: curvature (9.18b), outlier or large residual (9.18c), and high influence and leverage (9.18d). They are generally easy to identify from plots (y vs. x, or e vs. $\hat{y}$) in a linear regression with one explanatory variable. However, in multiple linear regression they are much more difficult to visualize or identify, requiring plots in multi-dimensional space. Thus numerical measures of their occurrence, called "regression diagnostics", have been developed.

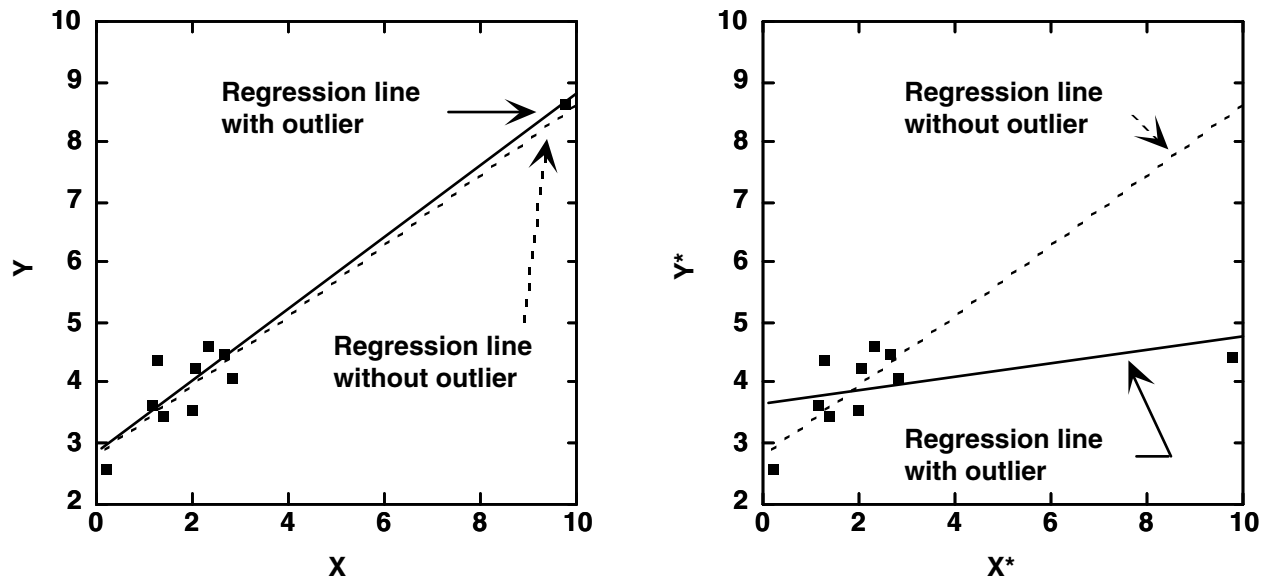
Equations for diagnostics useful in identifying points of leverage, influence, or outliers are given here in terms of the two dimensions (x,y) applicable to simple linear regression (SLR). Each can be generalized using matrix notation to a larger number of dimensions for multiple linear regression (MLR). Further references on regression diagnostics are Belsley, Kuh, and Welsch (1980), Draper and Smith (1981), and Montgomery and Peck (1982).

9.5.1 Measures of Outliers in the x Direction

9.5.1.1 Leverage

Leverage is a measure of an "outlier" in the x direction, as in graph 9.18a. It is a function of the distance from the ith x value to the middle (mean) of the x values used in the regression. Leverage is usually denoted as h_i, the ith diagonal term of the "hat" matrix $X(X'X)^{-1}X'$, or for SLR

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_x}.$$



a) outlier with high leverage, but low influence

b) outlier with high leverage and high influence

Figure 9.19 Influence of location of a single point on the regression slope.

A high leverage point is one where $h_i > 3p/n$ where p is the number of coefficients in the model (p=2 in SLR, b_0 and b_1). Though leverage is concerned only with the x direction, a high leverage point has the potential for exerting a strong influence on the regression slope. If the high leverage point falls far from the regression line that would be predicted if it were absent from the data set, then it is a point with high influence as well as high leverage (figure 9.19b).

9.5.2 Measures of Outliers in the y Direction

9.5.2.1 Standardized residual

One measure of outliers in the y direction is the standardized residual e_{si}. It is the actual residual $e_i = y_i - \hat{y}_i$ standardized by its standard error.

$$e_{si} = \frac{e_i}{s\sqrt{1 - h_i}}$$

An extreme outlier is one for which $|e_{si}|>3$. There should be only an average of 3 of these in 1,000 observations if the residuals are normally distributed. $|e_{si}|>2$ should occur about 5 times in 100 observations if normally distributed. More than this number indicates that the residuals do not have a normal distribution.

9.5.2.2 Prediction residuals and the PRESS statistic

A very useful form of residual computation is the prediction residual $e_{(i)}$. These are computed as $e_{(i)} = y_i - \hat{y}_{(i)}$ where $\hat{y}_{(i)}$ is the regression estimate of y_i based on a regression equation computed leaving out the ith observation. The (i) symbolizes that the ith observation is left out of the computation. These are easily calculated using leverage statistics without having to perform n separate regressions:

$$e_{(i)} = e_i / (1 - h_i) .$$

One of the best measures of the quality of a regression equation is the "PRESS" statistic, the "PRediction Error Sum of Squares."

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2$$

PRESS is a validation-type estimator of error. Instead of splitting the data set in half, one-half to develop the equation and the second to validate it, PRESS uses n−1 observations to develop the equation, then estimates the value of the one left out. It then changes the observation left out, and repeats the process for each observation. The prediction errors are squared and summed. Minimizing PRESS means that the equation produces the least error when making new predictions. In multiple regression it is a very useful estimate of the quality of possible regression models.

9.5.2.3 Studentized residuals

Studentized residuals (TRESIDs) are used as an alternate measure of outliers by some texts and computer software. They are often confused with standardized residuals.

$$\text{TRESID}_i = \frac{e_i}{s_{(i)}\sqrt{1-h_i}} = \frac{e_{(i)}\sqrt{1-h_i}}{s_{(i)}}$$

where

$$s^2_{(i)} = \frac{(n-p)\, s^2 - [\, e_{(i)}^2 / (1 - h_i)\,]}{n - p - 1}$$

TRESIDs are often similar to the standardized residuals e_{si}, but are computed using a variance $s^2_{(i)}$ which does not include their own observation. Therefore an unusually large observation does not inflate the estimate of variance used to determine whether that observation is unusual, and outliers are more easily detected. Under a correct model with normal residuals, TRESIDs have the theoretical advantage that they should follow a t-distribution with n−p−1 degrees of freedom.

9.5.3 Measures of Influence

Observations with high influence are those which have both high leverage and large outliers (figure 9.19b). These exert a stronger influence on the position of the regression line than other observations.

9.5.3.1 Cook's D

One of the most widely used measures of influence is "Cook's D" (Belsley et al., 1980).

$$D_i = \frac{e_i^2 h_i}{ps^2 (1 - h_i)^2} = \frac{e_{(i)}^2 h_i}{ps^2}$$

The ith observation is considered to have high influence if $D_i > F_{(p+1,n-p)}$ at α=0.1 where p is again the number of coefficients. Note that, for SLR with more than about 30 observations, the critical value for D_i would be about 2.4, and for MLR with several explanatory variables the critical value would be in the range of 1.6 to 2.0. Finding an observation with high Cook's D should lead to a very careful examination of the data value for possible errors or special conditions which might have prevailed at the time it occurred. If it can be shown that an error occurred, the point should be corrected if possible, or deleted if the error can't be corrected. If no error can be proven, two options can be considered. A more complex model which fits the point better is one option. The second option is to use a more robust procedure such as that based on Kendall's τ (for one x variable) or weighted least squares (for more than one x variable). These methods for "robust regression" are discussed in Chapter 10.

9.5.3.2 DFFITS

The second influence diagnostic, related to TRESIDs, is the DFFITS:

$$DFFITS_i = \frac{e_i \sqrt{h_i}}{s_{(i)} (1-h_i)} = \frac{e_{(i)} \sqrt{h_i}}{s_{(i)}}$$

An observation is considered to have high influence if $|DFFITS_i| \geq 2\sqrt{p/n}$.

The identification of outliers can be done with either standardized or studentized residuals, and the identification of highly influential points can be done with either DFFITS or Cook's D. The leverage statistic identifies observations unusual in x. PRESS residuals are rarely used except to sum into the PRESS statistic, in order to compare competing multiple regression models.

Example 1

The data of figure 9.19a were analyzed by regression, and the above diagnostics calculated. These data exhibit high leverage but low influence, as removal of the one outlier in the x direction will not appreciably alter the slope of the regression line. The regression results are given in Table 9.3. The only unusual value is the leverage statistic h_i for the last point, the one which plots to the right on the graph. A value of
$3p/n = 0.6$, so the 0.919 for this point shows it to be one of high leverage.

$y = 2.83 + 0.60\ x$

$n = 10$ $s = 0.43$ $R^2 = 0.94$

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	2.828	0.195	14.51	0.000
Slope β_1	0.596	0.054	10.98	0.000

OBS#	e_i	h_i	$e_{(i)}$	e std	e stud	DFFITS	D_i
1	−0.377	0.188	−0.465	−0.974	−0.970	−0.467	0.110
2	0.085	0.131	0.098	0.213	0.200	0.077	0.003
3	0.804	0.126	0.920	1.997	2.640	1.005	0.289
4	−0.219	0.122	−0.249	−0.543	−0.518	−0.193	0.020
5	−0.484	0.104	−0.541	−1.189	−1.226	−0.419	0.082
6	0.204	0.104	0.228	0.501	0.476	0.162	0.014
7	0.380	0.101	0.423	0.931	0.922	0.309	0.048
8	0.059	0.100	0.066	0.146	0.136	0.045	0.001
9	−0.462	0.101	−0.514	−1.132	−1.156	−0.388	0.072
10	0.010	**0.919**	0.132	0.087	0.081	0.276	0.043

Table 9.3 Regression statistics for the data of Figure 9.19a

Table 9.4 presents the analysis of the data for figure 9.19b. Note that the equation and ensuing R^2 are quite different. Only y for the 10th observation was changed from its previous value. Note also that the influence statistics DFFITS and D_i are large. The 10th observation is one of high influence, showing that the line computed with this point deleted is quite different than the one with it included. This is also demonstrated by the prediction residual $e_{(i)}$, whose absolute value is also large. The leverage statistic is unchanged from 9.19a, as the x position has not changed.

It is also quite important to note the values for the 10th observation which are not large -- the residual itself (e_i) and the standardized residual (e std). These statistics do not indicate the magnitude of the problem. Therefore residuals plots which use e_i or

e std may not display influential observations as such, because the line has been so drawn near to the outlier that its residual does not appear unusual.

9.5.4 Measures of Serial Correlation

One of the assumptions of regression is that the residuals e_i are independent. Many hydrologic data sets on which regression is performed are actually pairs of time series -- precipitation and flow, flow and concentration, concentration of one constituent versus concentration of another. These series often exhibit serial correlation, the dependence or correlation in time sequence between residuals, violating the assumption of independence (figure 9.10). If the sampling frequency is high enough, serial correlation of the residuals is virtually certain to exist. If serial correlation occurs, the following two problems ensue:

1) The estimates of the regression coefficients are no longer the most efficient estimates possible, though they remain unbiased, and
2) The value of s^2 may seriously underestimate the true σ^2.
   This means that all of the hypothesis tests are wrong (H_0 is rejected too easily) and that confidence and prediction intervals are too narrow.

$y^* = 3.65 + 0.11\ x^*$

$n = 10$ $s = 0.60$ $R^2 = 0.21$

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	3.648	0.270	13.53	0.000
Slope β_1	0.111	0.075	1.48	0.000

OBS#	e_i	h_i	$e_{(i)}$	e std	e stud	DFFITS	D_i
1	−1.096	0.188	−1.350	−2.042	−2.761	−1.330	0.483
2	−0.166	0.131	−0.192	−0.300	−0.282	−0.109	0.006
3	0.599	0.126	0.687	1.077	1.090	0.415	0.084
4	−0.370	0.122	−0.421	−0.663	−0.638	−0.238	0.030
5	−0.325	0.104	−0.363	−0.576	−0.551	−0.188	0.019
6	0.373	0.104	0.417	0.662	0.637	0.217	0.025
7	0.680	0.101	0.757	1.204	1.245	0.417	0.081
8	0.534	0.100	0.594	0.945	0.938	0.313	0.049
9	0.099	0.101	0.110	0.176	0.165	0.055	0.001
10	−0.329	**0.919**	**−4.117**	−1.955	−2.531	**−8.579**	**21.961**

Table 9.4 Regression statistics for the data of Figure 9.19b

One can search for the presence of serial correlation in two ways. The first is graphical: plotting e_i versus i or a measure of time (figure 9.10b). If there is a tendency for the data to "clump,"

positives follow positives, negatives follow negatives, this may mean there is dependence. The clumping could arise for four different reasons: long-term trend, seasonality, dependence on some other serially correlated variable which was not used in the model, serial dependence of residuals, or some combination of these. Examination of a graph of e_i versus time should help to reveal trend or seasonality if they exist. If there is reason to believe it is trend or seasonality (or both), then steps should be taken to remove these features from the residuals by adding additional explanatory variables. Similarly, if there is an important variable missing from the model, plots of e_i versus this variable should show it, and incorporating this new variable may remove the clumpiness of the residuals. This is particularly likely if this new explanatory variable exhibits serial dependence, seasonality, or trend. The residuals from these new regressions can be plotted again to see what effect this had.

9.5.4.1 Durbin-Watson statistic

There are also statistics for evaluating the dependence of residuals. The standard one is the Durbin Watson statistic (Durbin and Watson, 1951). It is very closely related to a serial correlation coefficient. The statistic is

$$d = \frac{\sum_{i=2}^{n} [e_i - e_{(i-1)}]^2}{\sum_{i=1}^{n} e_i^2}$$

A small value of d is an indication of serial dependence. The H_0 that the e_i are independent is rejected in favor of serial correlation when $d<d_L$ which is tabled in time-series texts. The value of d_L depends on the size of the data set, the number of explanatory variables, and α. However, a low value of d will not give any clue as to its cause. Thus, the graphical approach is vital, and the test is only a check. The Durbin Watson statistic requires data to be evenly spaced in time and with few missing values.

9.5.4.2 Serial correlation coefficient

Serial correlation can also be measured by the correlation coefficient between a data point and its adjacent point. As a linear relationship between pairs of points cannot be assumed, the Kendall's or Spearman's coefficients will provide robust measures of serial dependence. To compute whether this serial dependence is in fact significant,

1) Compute the regression between y and x.
2) Order the resulting residuals by the relevant time or space variable t_1 to t_n.

3) Offset or "lag" the vector of residuals to form a second vector, the lagged residuals. The residuals pairs then consist of (e_i, e_{i-1}) for all i from t_2 to t_n. Figure 9.10c plots one such set of data pairs, illustrating their correlation.
4) Compute Kendall's tau (or Spearman's rho) between the pairs (e_i, e_{i-1}). If the correlation is significant, the residuals are serially correlated.

9.5.4.3 What to do if serial correlation is present

If serial dependence cannot be removed by adding new variables, and one wants to make inferences about parameters, then these three options are available.

1) Sample from the data set. For example, if the data set is quite large and the data are closely spaced in time (say less than a few days apart), then simply discard some of the data in a regular pattern. The dependence that exists is an indication of considerable redundancy in the information, so not a great deal is lost in doing this.
2) Group the data into time periods (e.g., weeks, months) and compute a summary statistic for the period such as a time-weighted mean or median, a volume-weighted mean or median, and then use these summary statistics in the regression. This should only be done when the sampling frequency has remained unchanged over the entire period of analysis.
3) Use much more sophisticated estimation methods, specifically Box and Jenkins (1976) transfer function models, or regression with autoregressive errors Johnston (1984).

9.6 Transformations of the Response (y) Variable

The primary reason to transform the response variable is because the data are heteroscedastic -- the variance of the residuals is a function of x. This situation is very common in hydrology. For example, suppose a rating curve between stage (x) and discharge (y) at a stream gage has a standard error of 10 percent. This means that whatever the estimated discharge, the standard error is 10 percent of that value. The absolute magnitude of the variance around the regression line between discharge and stage therefore increases as estimated discharge increases. The ideal variance stabilizing transformation in these cases is the logarithm because a multiplicative relationship, such as standard error = 0.10•estimate, becomes a constant additive relationship after log transformation. This satisfies the regression assumptions. The two topics that require careful attention when transforming y are:

1) deciding if the transformation is appropriate, and
2) interpreting resulting estimates.

9.6.1 To Transform or Not to Transform?

The decision to transform y should generally be based on graphs. First develop the best possible non-transformed model. This should entail considering all sorts of transformations of x (or

multiple x variables) to get a good and reasonable fit. Then plot e_i vs. $\hat{y}_i$ to check for heteroscedasticity, do a probability plot for e_i to check for normality, and examine the function for unreasonable results (i.e., predictions of negative values for variables that can't go negative). If serious problems arise for any of these reasons, transform y and repeat the process. If both the transformed and untransformed scales have problems, then either look for a different transformation or accept the lesser of two evils.

Two methods are available to numerically judge whether or not to transform y. The first is to perform a series of transformations, perform regressions, and choose the transformation which maximizes the probability plot correlation coefficient (PPCC) for the regression residuals. This optimizes the normality of residuals. The second method is similar, optimizing for linearity. It searches for the minimum sum of squared errors SSE from a series of regressions using transformed and scaled y variables (Montgomery and Peck, 1982, p.94). The transformations used are scaled versions of the ladder of powers called "Box-Cox transformations". Scaling is required in order to compare the errors among models with differing units of y. Either numerical method can be a useful guide to selecting several candidate transformations from which to choose. However, the final choice should be made only after looking at residuals plots.

The key thing to note here is that **comparisons of R^2, s, or F statistics between transformed and untransformed models cannot easily be used to choose among them**. Each model is attempting to predict a different variable (y, log(y), 1/y, etc.). The above statistics therefore measure how well different variables are predicted, and so cannot be directly compared. Instead, the appropriate response variable is one which fits the assumptions of regression well -- linear and homoscedastic, having a good residuals plot. Once a hydrologist has developed some experience with certain kinds of data sets, it is quite reasonable to go directly to the appropriate transformation without a lot of investigation. One helpful generalization is that any y variable that covers more than an order of magnitude of values in the data set, as sediment discharge or bacterial densities typically do, probably needs to be transformed.

9.6.2 Consequences of Transformation of y

Let's take a particular, but rather common, case of a transformed regression problem. The model is

$$\ln(L) = \beta_0 + \beta_1 \ln Q + \varepsilon$$

where ln is the natural log, L is constituent load (tons/day), and Q is discharge (cubic feet per second). Let us further assume that the ε values are normal with mean zero and variance σ^2.

Figure 9.20 illustrates a data set typical of such L vs. Q data, shown here as a log-log plot. The lines results from a SLR done in log units. The middle line is the regression line and the 50% and 95% prediction intervals are shown. Note that, because of the normality assumption, the

prediction intervals are symmetric about the regression line. For any given Q value the five lines on the graph represent five different percentage points on the conditional distribution of ln(L). They are the 2.5, 25, 50 (median), 75, and 97.5 percentage points. The median also happens to be the conditional mean for ln(L) because when normality is assumed the median = mean. So the regression line falls on both the conditional median and mean value for ln(L).

Figure 9.21 takes each of these data points and lines and replots them in the original units (L versus Q). The five curves remain the 2.5, 25, 50, 75, and 97.5 percentage points on the conditional distribution. Now however this distribution of L conditional on Q is lognormal, not a normal distribution. Note the asymmetry of the curves around the regression line. For a lognormal distribution the mean is not equal to the median. While the central line remains the conditional median following transformation, the conditional mean of L will always lie somewhere above the regression line.

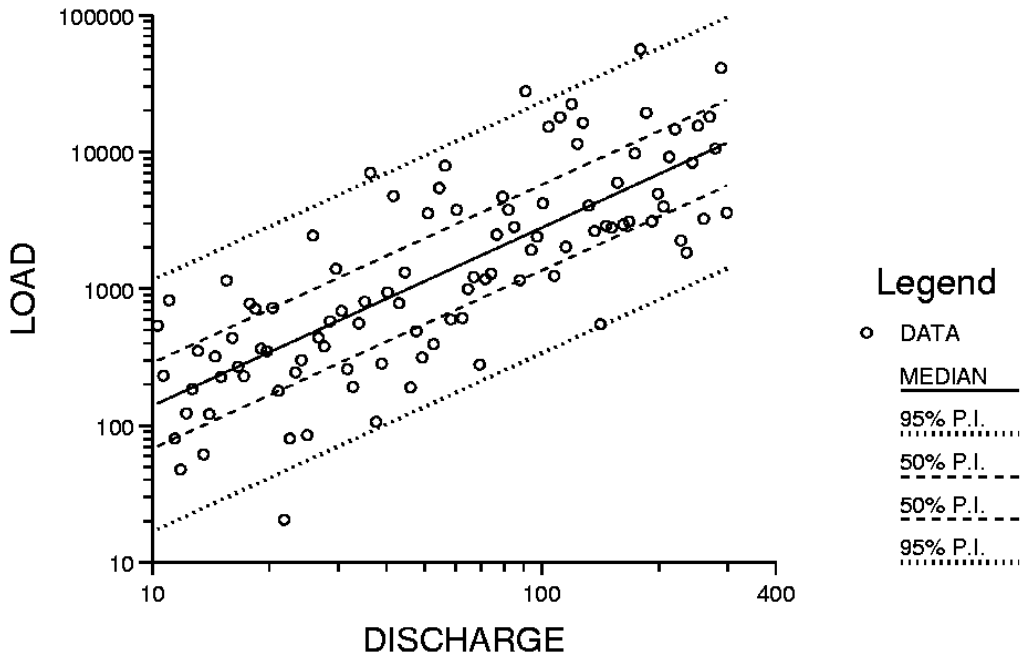


Figure 9.20 Prediction intervals and log-log regression in log units.

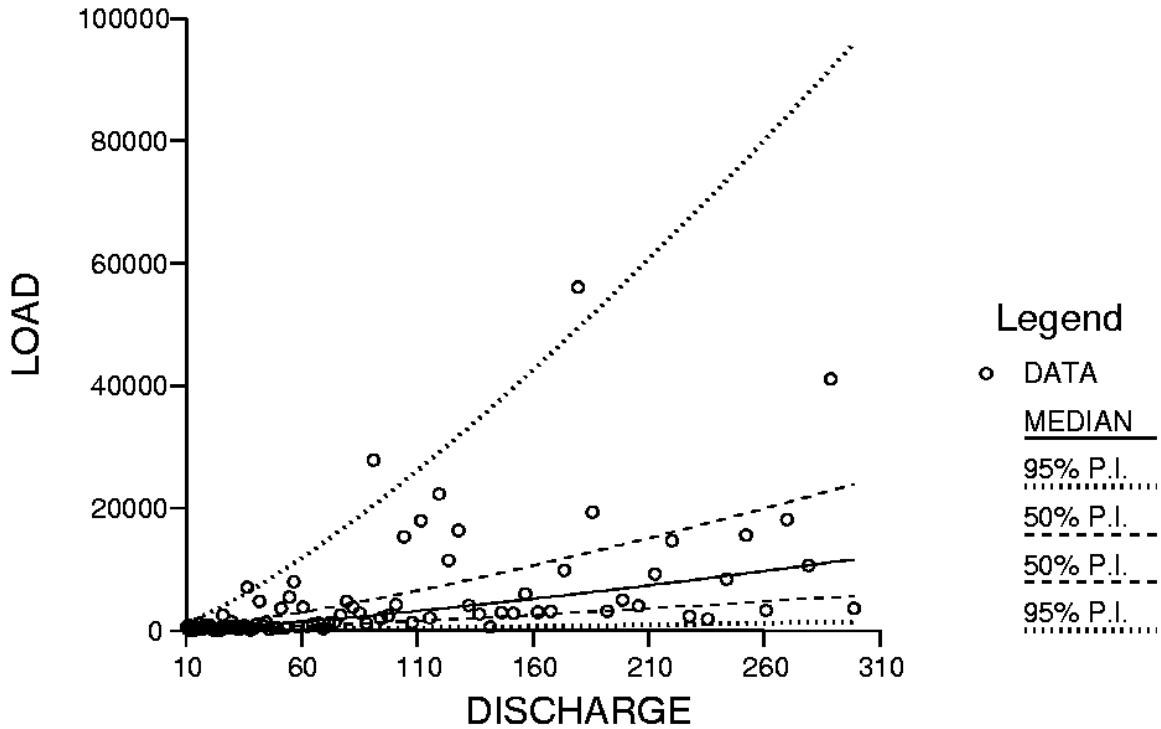


Figure 9.21 Prediction intervals and log-log regression re-expressed in original units.

9.6.3 Computing Predictions of Mass (Load)

9.6.3.1 Median or "rating curve" estimate of mass

When the objective is estimating the mass of sediment (or nutrient or contaminant) entering a lake, reservoir, or estuary, the mean for each of many short time periods can be estimated by regression and summed to estimate the total (or mean) mass over a longer time period. This is appropriate because the sum of the means equals the mean of the sum. However, simply transforming estimates from a log-regression equation back into the original units for y provides a median estimate of L, not a mean. Unfortuantely, this has been the traditionally-used method since Miller (1951). The sum of these medians provides an estimate of the mean of L which is biased low. As the sum of the medians is not the median of the sum, it is difficult to state what the sum of these median values represents, except that it underestimates the long-term mean load.

Ferguson (1986) points out for some very realistic cases that using the median or rating curve estimate for loads:

$$\hat{L}_m = \exp\,[b_0 + b_1 \ln(Q_0)]$$

will result in underestimates of the mean by as much as 50%. The question then is how to compensate for this bias. The following two methods, one assuming a normal distribution of the logs and the other a nonparametric method, attempt to correct for this bias of the median estimate.

9.6.3.2 Parametric or "MLE" estimate of mass

If the residuals in natural log units were known to be normal and the parameters of the model (β_0, β_1, σ^2) were known without error, the theory of the lognormal distribution (Aitchison and Brown, 1981) provides the following results:

$$\text{Median of L given } Q_0 = \exp[\beta_0 + \beta_1 \ln(Q_0)] = L_m = \exp[\beta_0] \bullet Q_0^{\beta_1}$$

$$\text{Mean of L given } Q_0 = E[L|Q_0] = \exp[\beta_0 + \beta_1 \ln(Q_0) + 0.5\,\sigma^2] = L_m \bullet \exp[0.5\,\sigma^2]$$

$$\text{Variance of L given } Q_0 = V[L|Q_0] = [L_m \bullet \exp(0.5\,\sigma^2)]^2 \bullet [\exp(\sigma^2) - 1]$$

These equations would differ if base 10 logarithms were used (Ferguson, 1986).

Unfortunately the true population values β_0, β_1, and σ^2 are never known in practice. All that is available are the estimates b_0, b_1, and s^2. Ferguson (1986) assumed these estimates were the true values for the parameters. His estimate of the mean is then

$$\hat{L}_{MLE} = \exp[b_0 + b_1 \ln(Q_0) + 0.5\,s^2]$$

When n is large (>30) and σ is small (<0.5), $\hat{L}_{MLE}$ is a very good approximation. However, when n is small or σ is large, it can <u>overestimate</u> the true mean -- it overcompensates for the bias. There is an exact unbiased solution to this problem which was developed by Bradu and Mundlak (1970). It is not given here due to the complexity of the formula. Its properties are discussed in Cohn (1988). Even so, the validity of Bradu and Mundlak's solution depends on the normality of the residuals which can never be assured in practice.

9.6.3.3 Nonparametric or "smearing" estimate of mass

There is an alternative approach which only requires the assumption that the residuals are independent and homoscedastic. They may follow any distribution. This is the "smearing" estimate of Duan (1983). In the case of the log transform it is

$$\hat{L}_D = \exp[b_0 + b_1 \ln(Q_0)] \bullet \frac{\sum_{i=1}^{n} \exp[e_i]}{n}$$

The smearing estimator is based on each of the residuals being equally likely, and "smears" their magnitudes in the original units across the range of x. This is done by re-expressing the residuals from the log-log equation into the original units, and computing their mean. This mean is the "bias-correction factor" to be multiplied by the median estimate for all x_0. Even when the residuals in log units are normal, the smearing estimate performs very nearly as well as Bradu and Mundlak's unbiased estimator. It avoids the overcompensation of Ferguson's approach. As it is robust to the distribution of residuals, it is the most generally-applicable approach.
The smearing estimator can also be generalized to any transformation. If Y = f(y) where y is the response variable in its original units and f is the transformation function (e.g., square root, inverse, or log), then

$$\hat{y}_D = \frac{\sum_{i=1}^{n} f^{-1}(b_0 + b_1 X_0 + e_i)}{n}$$

where b_0 and b_1 are the coefficients of the fitted regression and e_i are the residuals ($Y_i = b_0 + b_1 X_0 + e_i$) , f^{-1} is the inverse of the selected transformation (e.g., square, inverse, or exponential, respectively) and X_0 is the specific value of X for which we want to estimate y.

9.6.4 An Example

Total phosphorus loads are to be estimated for the Illinois River at Marseilles, Illinois, drainage area 8259 square miles, for the period 1972-1985. The data are contained in Appendix C10. The 96 measurements of load are plotted in figure 9.22 as a function of discharge. As loads were not sampled for each day during this time period, estimates of load for unsampled days are to be obtained from a regression equation as a function of discharge.

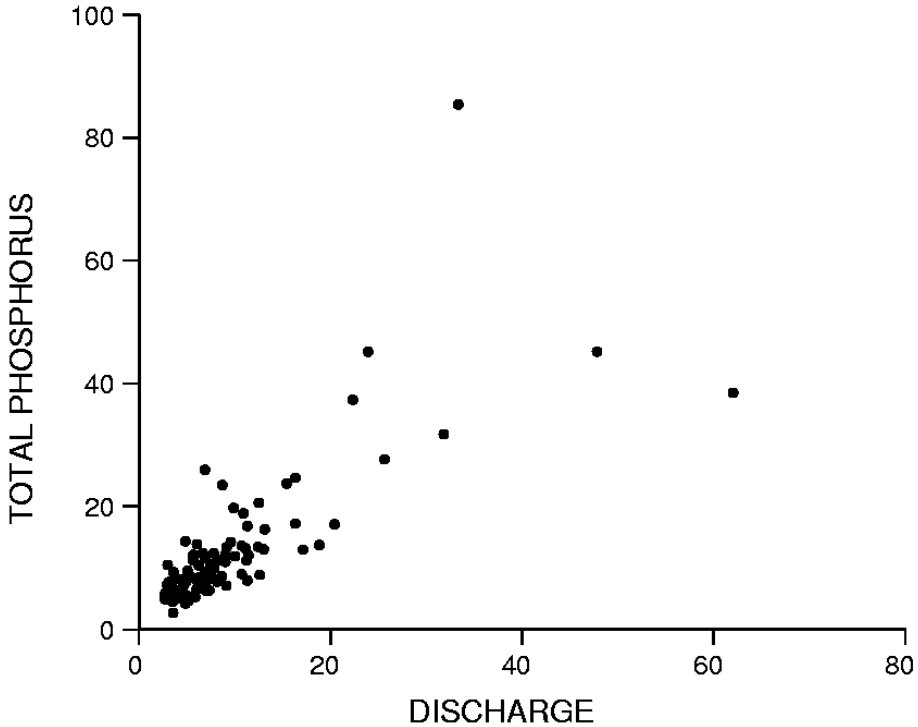


Figure 9.22 Total phosphorous load and stream discharge for the Illinois River

The first question is whether a log transform of load is necessary to develop a good prediction equation. From figure 9.22, the variance appears to greatly increase as discharge increases. Therefore a log transformation of phosphorus is attempted. This results in a curvilinear pattern, so the log of discharge is computed and used as the explanatory variable. As seen in figure 9.23, the transformation of both load and discharge results in a linear, homoscedastic relationship. A residuals plot in figure 9.24 shows little evidence of structure, indicating that the units are appropriate. Therefore these units are used for the regression. Table 9.5 gives the relevant regression statistics.

$\ln(L) = 0.80 + 0.76 \ln(Q)$

$n = 96$ $s = 0.339$ $R^2 = 0.68$

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	0.799	0.114	7.03	0.000
Slope β_1	0.761	0.054	14.10	0.000

Table 9.5 Regression statistics for the Illinois River phosphorus data

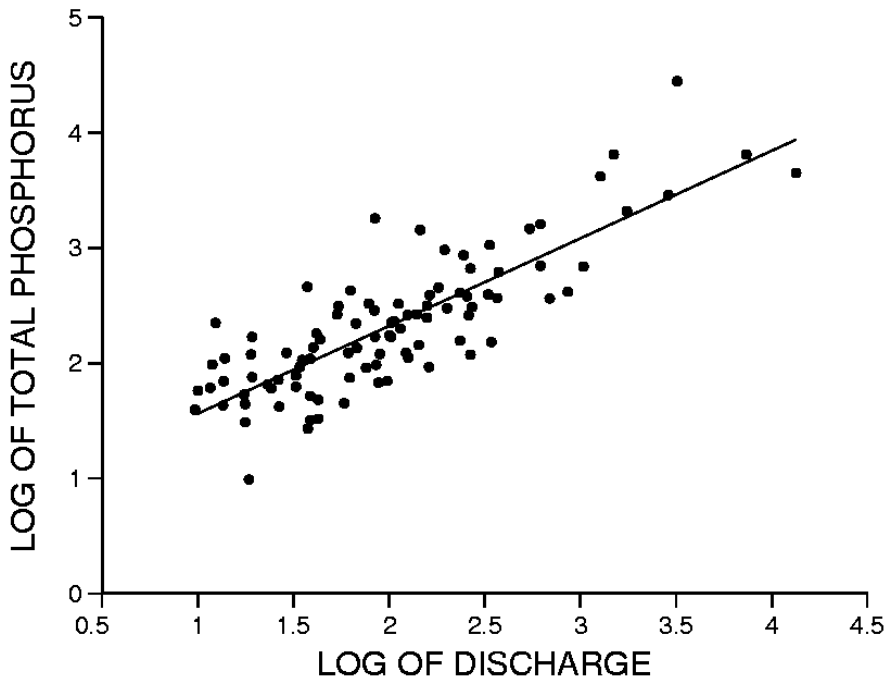


Figure 9.23 Log-log relation between phosphorous and discharge for the Illinois River

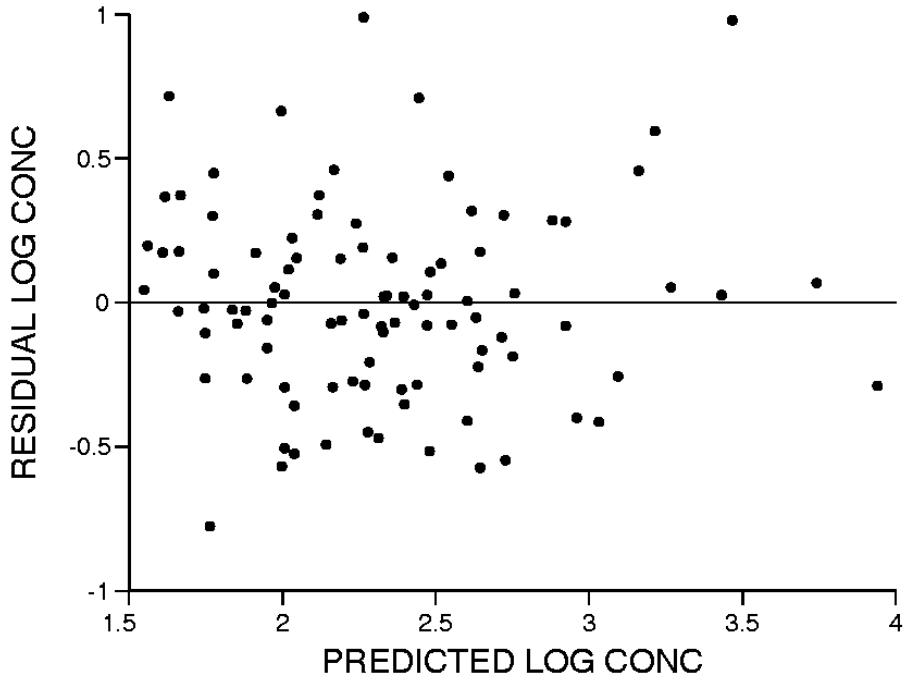


Figure 9.24 Residuals plot for ln(phosphorous) versus ln(discharge)

To illustrate the bias in phosphorus loads for the rating curve method, and the bias correction capabilities of the other two methods, estimates of all three will be computed here for the 96

days for which data exist. These values can then be compared to the "true" loads computed from the observed data.

The results from this regression are these	Mean Load	Error
true	= 12.64	--
median estimate	= 11.72	-7.3%
MLE estimate	= 12.41	-1.8%
smearing estimate	= 12.44	-1.6%

The median estimate is biased low, while the MLE and smearing estimates are close to each other and to the true value (figure 9.25). The MLE and smearing estimates should be expected to be similar here, as the residuals are fairly symmetric, n is large and s is small. These are the conditions under which the MLE works well. Had s been large (>1) or n small (<30) the MLE would probably have had a positive bias, and only the smearing estimate would have come close to the true value.

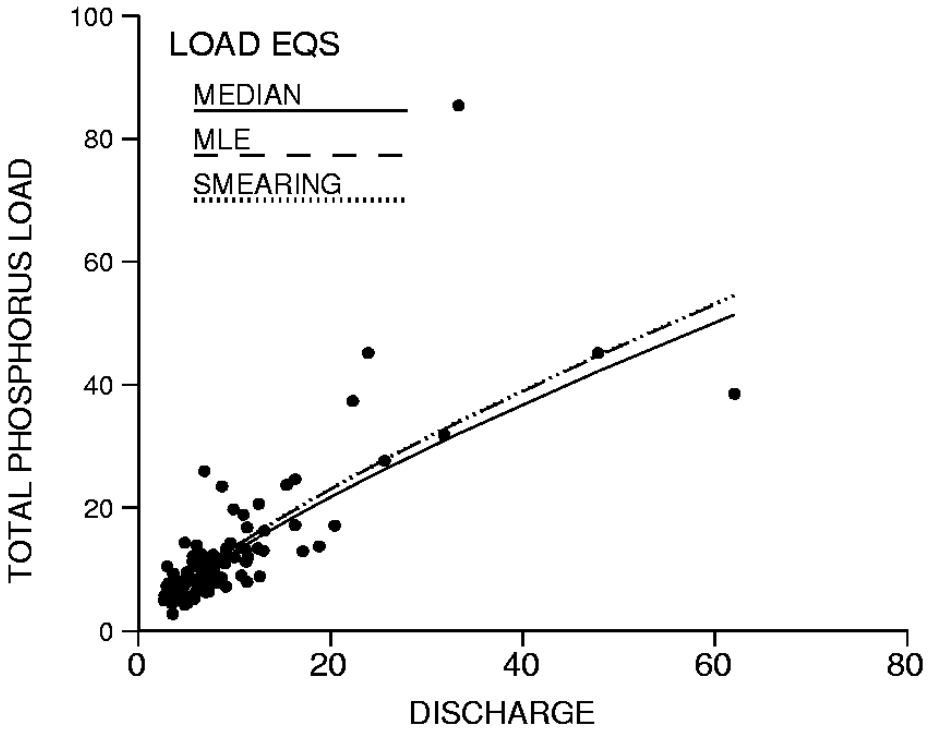


Figure 9.25 Load estimate curves with and without bias correction for Illinois R. data

9.7 Summary Guide to a Good SLR Model

1) "Should x be transformed, and if so, how?" Considerable help can come from a statistic such as R^2 (maximize it), or s (minimize it), but these numbers alone do not insure a good model. Many transformations can be rapidly checked with such statistics, but <u>always</u> look at a residual versus predicted plot before making a final decision. Transform x if the residuals plot appears non-linear but constant in variance, always striving for a linear relation between y and x.

2) "Should y be transformed, and if so, how?" Visually compare the transformed-y model to the untransformed-y model using their residuals plots (residual versus predicted). The better model will be more:
   1) linear,
   2) homoscedastic, and
   3) normal in its residuals.

   The statistics R^2, s, t-statistics on β_1, etc. will <u>not</u> provide correct information for deciding if a transformation of y is required.

   Should estimates of mass (loads) be developed using an equation having transformed-y units, the transformation bias inherent in the process must be compensated for by use of the smearing estimate, or MLE estimate when appropriate.

When there are multiple explanatory variables, more guidelines are required to choose between the many possible combinations of adding, deleting and transforming the various x variables. These are discussed in Chapter 11.

Exercises

9.1 Bedinger (1961) graphically related median grain size of alluvial aquifer materials in the Arkansas River Valley to their yield, in gallons per day per square foot. This enabled estimates of yield to be made at other locations based on measured grain-size analyses. Compute a regression equation to predict yield, based on the data in Appendix C11.

9.2 Estimate the mean yield in gallons per day per square foot available from four wells which together compose the public supply of a small town in the Arkansas River Valley. The wells have screens with identical cross-sectional areas. Median grain sizes for the units they draw from are: 0.1, 0.2, 0.4 and 0.6 millimeters.

9.3 Find a transformation of discharge for the Cuyahoga River TDS example which might improve on the $\log_{10}$ transformation used throughout the chapter. The data are found in Appendix C9. Obvious candidates include the ladder of power transformations. Another class of transformations that has been shown to work well for surface-water chemistry is the hyperbolic transformations (see Johnson, et al., 1969). The form of this transformation is $x=1/(1+kQ)$ where k is some constant supplied by the hydrologist. Some general advice about selecting k is that it's not worth the effort to try and get it "right" to a precision better than about half an order of magnitude. A good range to work in is

$$1/(100(\bar{Q}) < k < 100/(\bar{Q})$$

where $(\bar{Q})$ is the mean discharge.

The questions you should answer are:

a) What is a good transformation of Q to use in estimating TDS? (There is no "best" transformation, but there are several good ones.)
b) Describe your preferred model and indicate some reasons you might be concerned about it and might want to take steps to "fix" it in some fashion. (You will get a chance to later.)
c) What does it tell you about TDS behavior in the Cuyahoga River?
d) A question for the mathematically inclined. If k is set to some very large value (say around $100/\bar{Q}$), what other model does the hyperbolic approximate? If b is set to some very small value (say around $1/100\bar{Q}$), what other model does it approximate?

9.4 Objections have been raised to regressions such as load (L) versus stream discharge (Q) because Q is used to calculate L. This "spurious correlation" between Q and L can be avoided by using concentration (C) instead of load as the dependent variable. Loads would then be predicted from the estimated C. What do you think? How will the results using C compare to those using L as the regression's response variable? To answer this, perform the regression for the Illinois phosphorus data of section 9.6.4 and produce the 96 load estimates using ln(C in mg/L) instead of ln(L in tons per day). The data are found in Appendix C10. Note that the units of Q (thousands of cfs) mean that L = 2.7 Q•C. What happens to the regression coefficients and the associated statistics such as R^2, s, t-ratios, etc., when ln(C) rather than ln(L) is used? What is the appropriate conclusion to this controversy?

Chapter 10

Alternative Methods for Regression

Concentrations appear linearly related to distance down-dip in an aquifer. OLS regression shows the residuals to be of generally constant variance. However, several outliers in the data set inflate the standard error, and what appears graphically as a strong linear relationship tests as being insignificant due to the outliers' influence. How can a more robust linear fit be obtained which is not overly sensitive to a few outliers, and describes the linear relation between concentration and distance?

A water supply intake is to be located in a stream so that water elevation (stage) is below the intake only 5 percent of the time. Monitoring at the station is relatively recent, so OLS relating this and a nearby site having a 50 year record is used to generate a pseudo 50-year stage record for the intake station. The 5th percentile of the pseudo record is used as the intake elevation. Given that OLS estimates are reduced in variance compared to actual data, this elevation estimate will not be as extreme as it should be. What alternatives to OLS would provide better estimates?

The mass of a radionuclide present within the aquifer of one county was computed by performing a regression of concentration versus log of the hydraulic conductivity measured at 20 wells. This equation was used to generate estimates at 100 locations of known hydraulic conductivity, which are then multiplied by the volumes of water, and summed. However, the regression equation shows a marked increase in variance of concentration with increasing conductivity, even though the relationship is linear. Transformations may produce a nonlinear relationship, with probable transformation bias. An alternative to OLS is therefore required to account for heteroscedasticity without employing a transformation.

Situations such as the above frequently arise where the assumptions of constant variance and normality of residuals required by OLS regression are not satisfied, and transformations to remedy this are either not possible, or not desirable. In addition, the inherent reduction in variance of OLS estimates is not appropriate when extending records. In these situations, alternative methods are better for fitting lines to data. These include nonparametric rank-based methods, lines which minimize other than the squared residuals, and smooths.

10.1 Kendall-Theil Robust Line

The significance of a linear dependence between two continuous variables Y and X or their transforms may be tested by determining whether the regression slope coefficient for the explanatory variable is significantly different from zero. This is equivalent to the test for significance of the linear correlation coefficient r between Y and X. In a similar fashion, Kendall's rank correlation coefficient tau (see Chapter 8) may be used to test for any monotonic, not just linear, dependence of Y on X. Related to tau is a robust nonparametric line applicable when Y is linearly related to X. This line will not depend on the normality of residuals for validity of significance tests, and will not be strongly affected by outliers, in contrast to OLS regression.

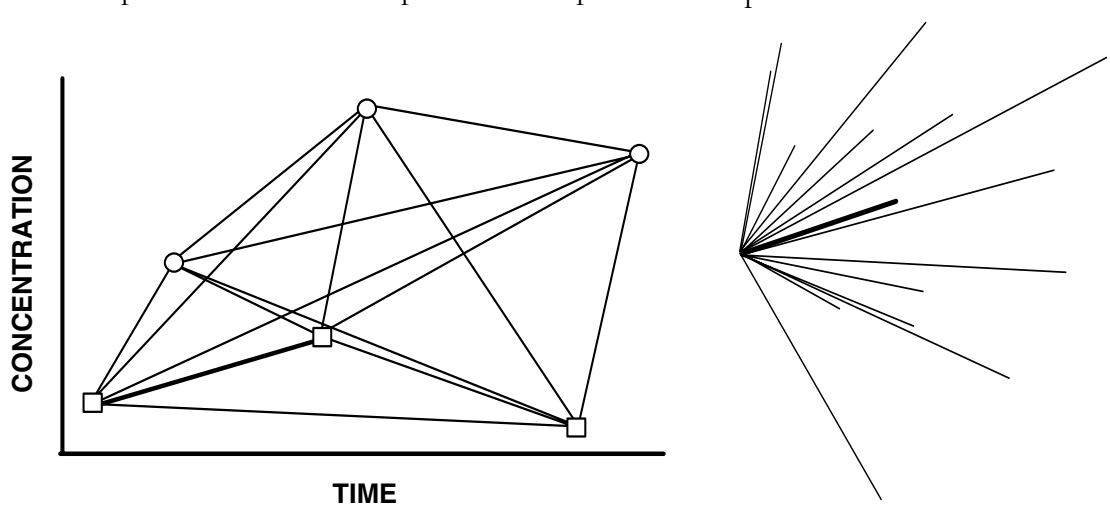
The robust estimate of slope for this nonparametric fitted line was first described by Theil (1950). An estimate of intercept is also available (Conover, 1980, p. 267). Together these define an estimate of a complete linear equation of the form:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 \bullet X$$

This line is closely related to Kendall's tau, in that the significance of the test for H_0: slope $\beta_1=0$ is identical to the test for H_0: tau=0.

10.1.1 Computation of the Line

The Theil slope estimate $\hat{b}_1$ is computed by comparing each data pair to all others in a pairwise fashion. A data set of n (X,Y) pairs will result in n(n−1)/2 pairwise comparisons. For each of these comparisons a slope $\Delta Y/\Delta X$ is computed (figure 10.1). The median of all possible pairwise slopes is taken as the nonparametric slope estimate $\hat{b}_1$.



A. B.

Figure 10.1 A. All possible pairwise slopes between six data points.
B. All possible slopes rearranged to meet at a common origin
The thick line is the median of the 15 slopes.

$$\hat{b}_1 = \text{median}\ \frac{(Y_j - Y_i)}{(X_j - X_i)} \qquad \text{for all } i < j \text{ and } i=1,2,...(n-1) \quad j=2,3,...n. \qquad [10.1]$$

<u>Example 1</u>

For example, given the following seven (X,Y) data pairs:

Y:	1	2	3	4	5	16	7
X:	1	2	3	4	5	6	7
Slopes:	+1	+1	+1	+1	+11	–9	
	+1	+1	+1	+6	+1		
	+1	+1	+4.3	+1			
	+1	+3.5	+1				
	+3	+1					
	+1						

There are (7)(6)/2 = 21 pair-wise slopes. Comparing points 2 and 1, slope = +1. Going down the column under point 1, comparing points 3 and 1, slope = +1.

For points 4 and 5 vs 1, slopes = +1. Comparing points 6 and 1, slope = (15/5) = +3, etc. After computing all possible slopes, they are put into ascending order:

–9, +1, +1, +1 +1 +1 +1 +1 +1 +1 +1
+1 +1 +1 +1 +1 +3 +3.5 +4.3 +6 +11

The median of these 21 values is the 11th smallest, or +1, so that $\hat{b}_1 = +1$.

The intercept is defined as follows

$$\hat{b}_0 = Y_{med} - \hat{b}_1 \bullet X_{med} \qquad [10.2]$$

where X_{med} and Y_{med} are the medians of X and Y respectively. This formula assures that the fitted line goes through the point (X_{med} ,Y_{med}). This is analogous to OLS, where the fitted line always goes through the point ($\bar{x}$,$\bar{y}$). For the Example 1 data above, X_{med} = 4 and Y_{med}= 4, so that $\hat{b}_0 = 4 - 1 \bullet 4 = 0$.

Other estimates of intercept have been suggested. One is the median of all possible intercepts computed by solving the Kendall line using $\hat{b}_1$ and each data point (Dietz, 1989). However, the estimate of intercept produced by placing the line through the data medians was found by Dietz to be efficient in the presence of outliers and non-normal residuals, while also being simple to compute. It is the estimate recommended here, due to its robustness and efficiency, simplicity, and analogy to OLS.

10.1.2 Properties of the Estimator

OLS regression for the example 1 data would produce a slope b_1 of 1.71. This differs substantially from the Theil estimate $\hat{b}_1$ of 1, due to the strong effect on the regression slope of the one outlying Y value of 16. This effect can be seen by changing the 6th Y value from 16 to 6. The regression slope would change from 1.71 to 1, but $\hat{b}_1$ would be unchanged. Similarly, if

the data value were changed from 16 to 200, b_1 would be greatly inflated while $\hat{b}_1$ would again remain at 1. The estimator $\hat{b}_1$ is clearly resistant to outliers. It responds to the bulk of the data.

$\hat{b}_1$ is an unbiased estimator of the slope of a linear relationship, and b_1 from OLS is also an unbiased estimator. However, the variance of the estimators differ. When the departures from the true linear relationship (true residuals) are normally distributed, OLS is slightly more efficient (has lower variance) than the Kendall-based line. When residuals depart from normality (are skewed or prone to outliers), then $\hat{b}_1$ can be much more efficient than the OLS slope. The efficiency of the Theil estimate to the OLS slope is the same as that for the Hodges-Lehmann estimator in comparison to the mean (Sen, 1968), as the Theil estimate is one of the class of Hodges-Lehmann estimators. The Kendall-Theil line has the desirable properties of a nonparametric estimator: almost as "good" (efficient) as the parametric estimator when all assumptions of normality are met, and much better when those assumptions are not met.

One commonly-asked question is "how much of a departure from a normal distribution is necessary before a nonparametric test has an advantage over its parametric counterpart?". In the case of the Theil and OLS slope estimates, how non-normal must residuals be before the Theil estimate should be used? Are there advantages even in cases where the departure from normality is so small that visual inspection of the data distribution, or formal tests of normality, are unlikely to provide evidence for the lack of normality? Hirsch et al. (1991) tested the two slope estimators under one type of departure from normality, a mixture of two normal distributions. The predominant distribution had a mean of 10 and a standard deviation of 1; the second distribution had a mean of 11 and a standard deviation of 3. Figure 10.2 displays the two individual distributions and figure 10.3 displays a mixture of 95 percent from the first distribution and 5 percent from the second. Visual examination of figure 10.3 reveals only the slightest departure from symmetry. Given sampling variability that would exist in an actual data set it would be exceedingly unlikely that samples from this distribution would be identified as non-normal. Figure 10.4 displays a more substantial departure from normality, a mixture of 80 percent of the first distribution and 20 percent of the second. There is a difference in the shape of the two tails of the distribution, but again the non-normality is not highly noticeable.

Random samples were generated from each of several different mixture distributions containing between 0 and 20 percent of the second distribution. Data from each mixture were treated as a separate response variable in a regression versus a random order x. The true population slope is therefore zero. Both OLS and the Theil slope estimators were computed, and their standard deviations around zero recorded as root mean square error (RMSE). The results are given in figure 10.5 as the ratio of RMSE for the Theil estimator to the RMSE of the regression estimator (Hirsch et al., 1991). A value larger than 1 shows an advantage to OLS; smaller than 1 indicates the Theil estimate to be superior. For the larger sample size (n=36) the OLS estimator was more efficient (by less than 10 percent) when the data are not mixed and

therefore normal. With even small amounts of mixtures the Theil estimator quickly becomes more efficient. At a 20 percent mixture the Theil estimator was almost 20 percent more efficient. When the sample size was very small (n=6, smaller than typically used in a case study), efficiencies of the two methods were virtually identical.

These results reinforce that when the data or their transforms exhibit a linear pattern, constant variance and near-normality of residuals, the two methods will give nearly identical results. The advantages of familiarity and availability of diagnostics, etc. favor using OLS regression. However, when residuals are not normally distributed, and especially when they contain outliers, the Kendall method will produce a line with greater efficiency (lower variability and bias) than does OLS. Only small departures from normality (not always sufficient to detect with a test or histogram of residuals) favor using a robust approach. Certainly one should check all outliers for error, as discussed in Chapter 1. Do these represent a condition different from the rest of the data? If so, they may be the most important points in the data set. Perhaps another transformation will make the data more linear and residuals near-normal. But outliers cannot automatically be deleted, and often no error can be found. Robust methods like Kendalls or weighted least squares (discussed in sections 10.3 and 10.4) provide protection against disproportionate influence by these distinctive, but perhaps perfectly valid, data points.

For analysis of a small number of data sets, detailed searches for transformations to meet the assumptions of OLS are feasible. OLS is particularly informative in more complex applications requiring incorporation of exogenous effects using multiple regression (see Chapter 11). Cases aren't unusual, however, where no power transformations can produce near-normality due to heavy tails of the distribution. Perhaps the two greatest uses for Kendall's robust fit are 1) in a large study where multiple variables are tested for linear fits at multiple locations without the capability for exhaustive checking of distributional assumptions or evaluations of the sensitivity of results to outliers, and 2) by practitioners not trained in residuals plots and use of transformations to stabilize skewness and heteroscedasticity. A third use is for fitting lines to data which one does not wish to transform.

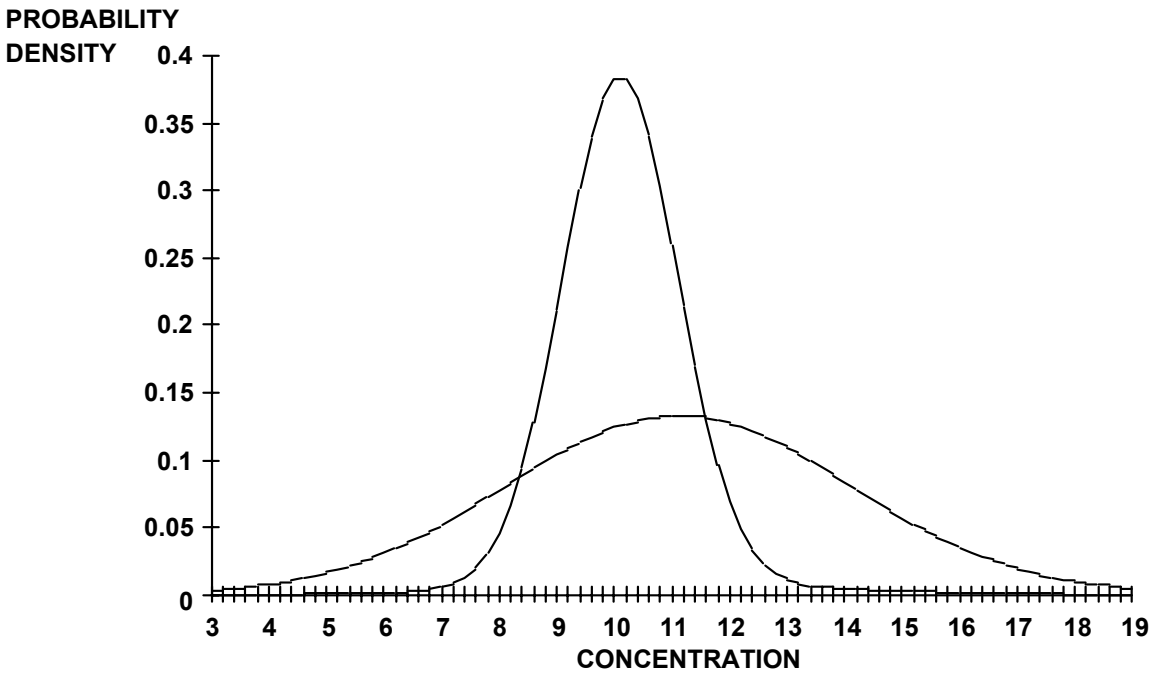


Figure 10.2. Two normal distributions, the first with mean = 10 and standard deviation = 1; the second with mean = 11 and standard deviation = 3 (from Hirsch et al., 1991).

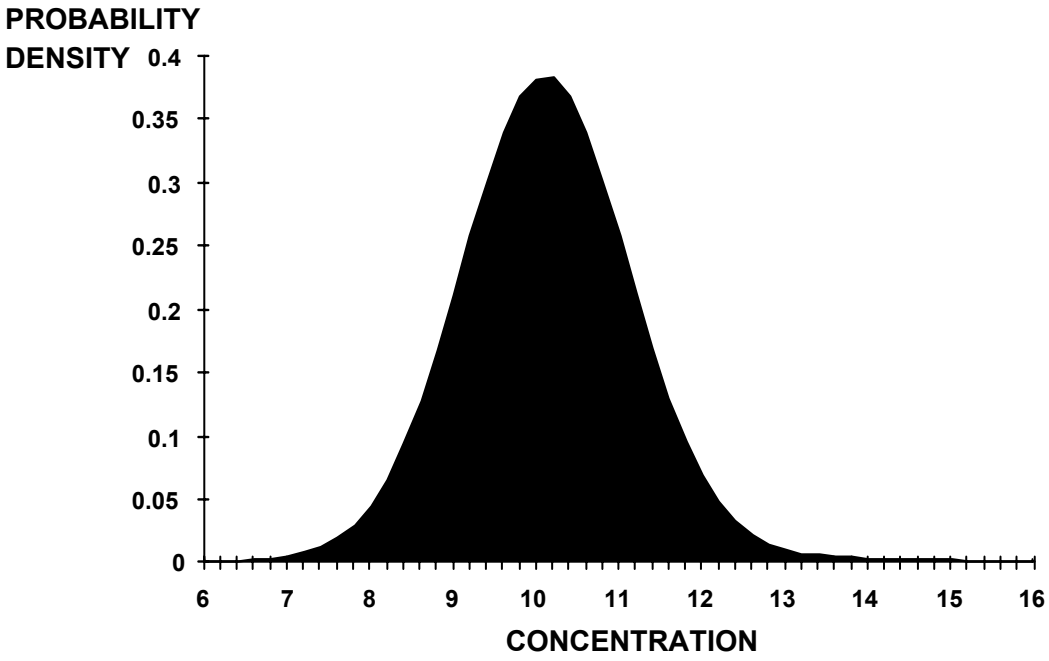


Figure 10.3. A mixture of data from distribution 1 (95 percent) and distribution 2 (5 percent) shown in figure 10.2 (from Hirsch et al., 1991).

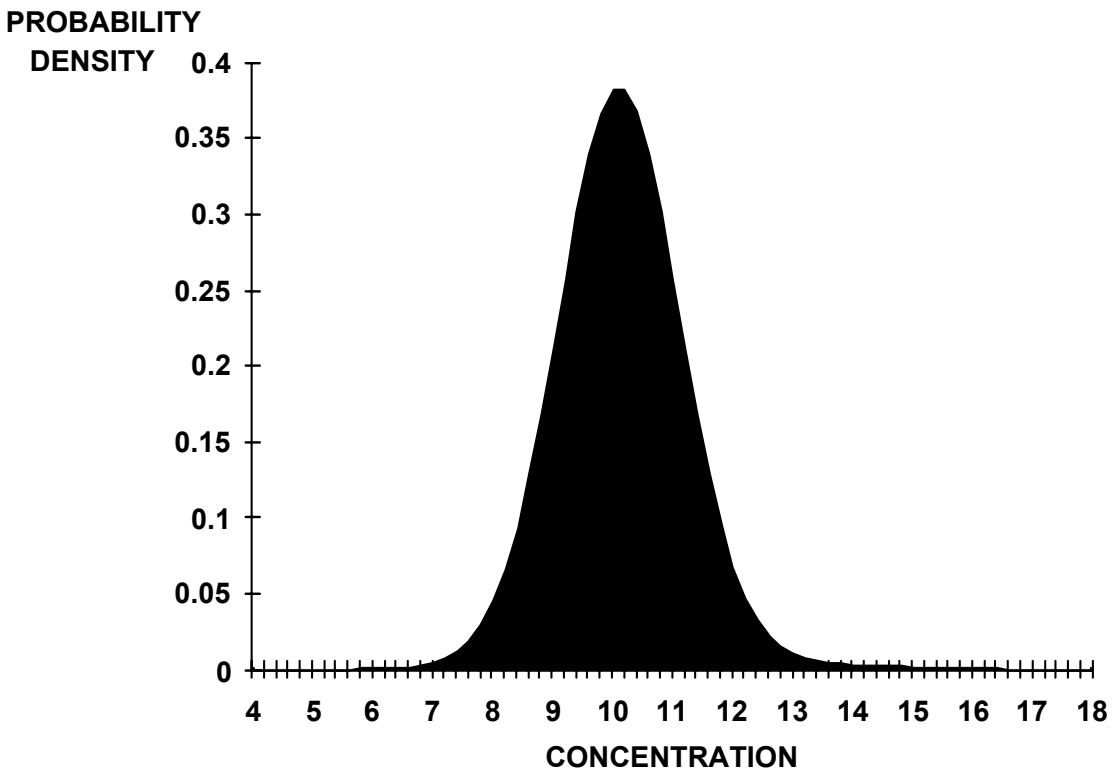


Figure 10.4. A mixture of data from distribution 1 (80 percent) and from distribution 2 (20 percent) shown in figure 10.2 (from Hirsch et al., 1991).

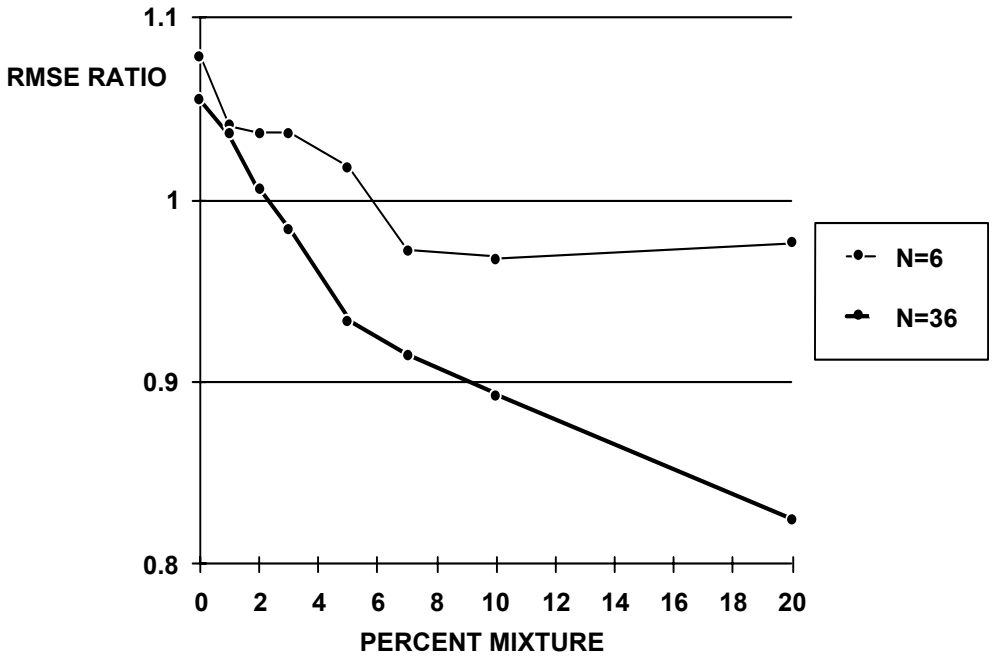


Figure 10.5. Relative efficiency of the Theil slope estimator as compared with the OLS slope. Efficiency is the ratio of the Theil RMSE to the OLS RMSE, expressed as a function of population mixture and record length (from Hirsch et al., 1991).

Example 2

Figure 10.6 shows an OLS and Kendall-Theil fit to trends in total phosphorus concentrations from 1975 to 1989 in the St. Louis R. at Scanlon, MN. The outliers are accurate values from floods, and therefore cannot be ignored or deleted. The question is whether there is a significant linear trend in concentration over this 14 year period. Here linear fits of concentration versus time are used to test for trend (see Chapter 12 for more on trend tests). The OLS slope is affected by the outliers present. Although the magnitude of the OLS estimate is similar to the Theil slope, the OLS slope does not test as significantly different from zero (p=0.43). This is due to inflation of the standard error by outliers in violation of the assumed normality of residuals. The Theil slope is highly significantly different from zero (p<0.0001). The Kendall-Theil line is not dependent on assumptions of normality which the data strongly violate.

10.1.3 Test of Significance

The test for significance of the Kendall-Theil linear relationship is the test for H_0: $\tau = 0$. This involves computation of Kendall's S statistic (equation 8.1 of Chapter 8). For n>10, the large sample approximation (equation 8.3 of Chapter 8) may be used. The Theil slope estimator $\hat{b}_1$ is closely related to Kendall's S and τ in the following ways.

1. S is the sum of the algebraic signs of the possible pairwise slopes.
2. If the amount ($\hat{b}_1$ X) is subtracted from every Y value, the new Y values will have an S and τ very close to zero, indicating no correlation.

If X is a measure of time, as it is for a trend test, subtracting ($\hat{b}_1$ X) yields a trend-free version of the Y data set.

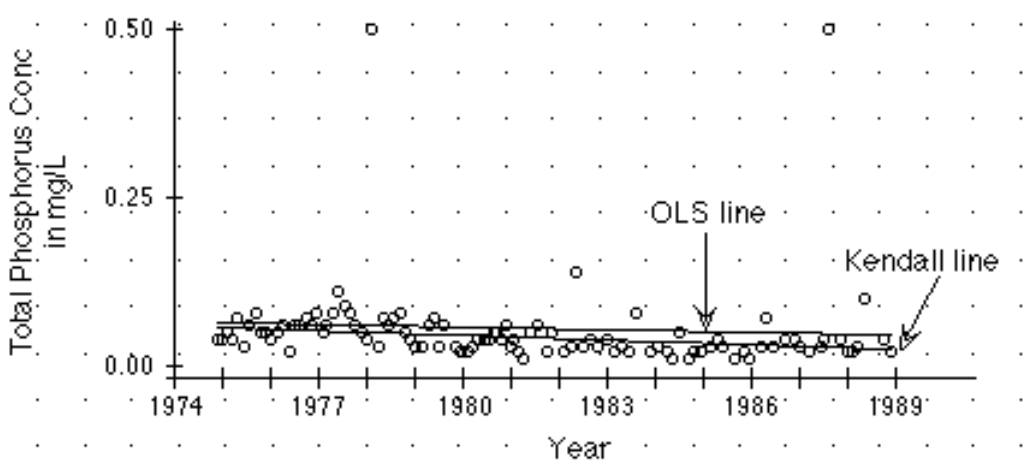


Figure 10.6. Total phosphorus concentrations with OLS and Kendall-Theil fitted lines for the St. Louis River at Scanlon, MN, 1975-1989.

Example 1, cont.

For the example 1 data set, the test of significance is computed as follows. S equals the sum of the signs of pairwise slopes already computed. There are n(n–1)/2 =21 slopes, 20 of which are positive and 1 negative, so that S = 20–1 = 19. Tau = 19/21 = 0.90. Using table B8 of the Appendix due to the small sample size, the exact two-sided p-value for an S of 19 and n=7 is 2•0.0014 = 0.003. (Inappropriately using the large sample approximation for such a small data set, the approximate p-value is 0.007.) Thus Y is significantly related to X in a linear fashion.

10.1.4 Confidence Interval for Theil Slope

Confidence intervals may be computed for the Theil slope $\hat{b}_1$ with procedures parallel to those used for other Hodges-Lehmann type estimators of earlier chapters. As before, the tabled distribution of the test statistic, in this case table B8 for the exact Kendall's test statistic or a table of standard normal quantiles for the large-sample approximation, is entered to find upper and lower limits corresponding to critical values at one-half the desired alpha level. These critical values are transformed into the ranks corresponding to data points at the ends of the confidence interval.

For small sample sizes, table B8 is entered to find the critical value X_u having a p-value nearest to $\alpha/2$. This critical value is then used to compute the ranks R_u and R_l corresponding to the slope values at the upper and lower confidence limits for $\hat{b}_1$. These limits are the R_lth ranked data points going in from either end of the sorted list of N = n•(n–1)/2 pairwise slopes. The resulting confidence interval will reflect the shape (skewed or symmetric) of the original data.

$$R_u = \frac{(N + X_u)}{2} \qquad [10.3]$$

$$R_l = \frac{(N - X_u)}{2} + 1 \qquad [10.4]$$

Example 1, cont.

The N=21 possible pairwise slopes between the n=7 data pairs for example 1 were:

–9,	+1,	+1,	+1	+1	+1	+1	+1	+1	+1	+1
+1	+1	+1	+1	+1	+3	+3.5	+4.3	+6	+11.	

$\hat{b}_1$ was the median or 11th largest slope. To determine a confidence interval for $\hat{b}_1$ with $\alpha \cong 0.05$, the tabled critical value X_u nearest to $\alpha/2$= 0.025 is found to be 15 (p=0.015). The rank R_u of the pairwise slope corresponding to the upper confidence limit is therefore

$$R_u = \frac{(21 + 15)}{2} = 18 \qquad \text{for } N=21 \text{ and } X_u=15.$$

The rank R_l of the pairwise slope corresponding to the lower confidence limit is

$$R_l = \frac{(21 - 15)}{2} + 1 = 4.$$

So an $\alpha = 2 \bullet 0.015 = 0.03$ confidence limit for $\hat{b}_1$ is the interval between the 4th and 18th ranked pairwise slope (the 4th slope in from either end), or

$$+1 \leq \hat{b}_1 \leq +3.5 .$$

The asymmetry around the estimate $\hat{b}_1 = 1$ reflects the low probability that the slope is less than 1, based on the data.

When the large-sample approximation is used, the critical value $z_{\alpha/2}$ from a table of standard normal quantiles determines the upper and lower ranks of the pairwise slopes corresponding to the ends of the confidence interval. Those ranks are

$$R_u = \frac{N + z_{\alpha/2}\sqrt{\frac{n(n-1)(2n+5)}{18}}}{2} + 1 \qquad [10.5]$$

$$R_l = \frac{N - z_{\alpha/2}\sqrt{\frac{n(n-1)(2n+5)}{18}}}{2} \qquad [10.6]$$

As an example, for n=20 pairs of data there would be N=(20)(19)/2 = 190 possible pairwise slopes. $\hat{b}_1$ is the average of the 95th and 96th ranked slopes. For a 95 percent confidence interval on $\hat{b}_1$, $z_{\alpha/2} = 1.96$ and

$$R_u = \frac{190 + 1.96 \bullet \sqrt{950}}{2} + 1 \qquad = 126.2$$

$$R_l = \frac{190 - 1.96 \bullet \sqrt{950}}{2} \qquad = 64.8$$

the 64.8th ranked slope from either end. Rounding to the nearest integer, the 126th and 65th ranked slopes are used as the ends of the α=0.05 confidence limit on $\hat{b}_1$.

Further discussion of these equations is in Hollander and Wolfe (1973), pp. 207-208.

10.2 Alternative Parametric Linear Equations

Hirsch and Gilroy (1984) described additional methods for fitting straight lines to data whose slopes and intercepts are computed using moment statistics. These lines differ from the OLS line of Chapter 9, and are more appropriate than that line for certain situations. For example, when X is to be predicted from Y using OLS, the resulting line differs from the OLS line predicting Y from X. This has implications for calibration. When many predictions are to be made and the distribution of those predictions is important (percentiles or spreads are of interest, as well as the mean), the Line of Organic Correlation (LOC) should be used instead of OLS. When describing a functional relationship between two variables without trying to predict

one from the other, LOC is again more appropriate than OLS. When some geographic trajectory is to be computed, the Least Normal Squares (LNS) line should be used.

10.2.1 OLS of X on Y

The OLS regression of Chapter 9 considered the situation where a response variable Y was to be modeled, enabling estimates of Y to be predicted from values of an explanatory variable X. Estimates of slope and intercept for the equation were obtained by minimizing the sum of squares of residuals in units of Y. Thus its purpose was to minimize errors in the Y direction only, without regard to errors in the X direction. The equation may be written as:

$$Y_i = \bar{Y} + r\,\frac{s_y}{s_x}\,(X_i - \bar{X}) \qquad [10.7]$$

where r is Pearson's linear correlation coefficient, s_y and s_x are the standard deviations of the Y and X variables, and $(r\, s_y/s_x) = (r\sqrt{SS_y}/\sqrt{SS_x}) = b_1$, the OLS estimate of slope (see Chapter 9). Assuming the linear form of the model is correct and that X and Y are measured without error, OLS will lead to estimates of Y_i for any given X_i which are unbiased and have minimum variance. This means that OLS is the preferred method of estimating a single value of Y given a value of X, where X is measured without error.

In contrast, situations occur where it is just as likely that X should be predicted from Y, or that the two variables are equivalent in function. One classic example is in geomorphology, where relations between the depth and width of a stream channel are to be related. It is as reasonable to perform a regression of depth on width as it is of width on depth. A second example is the relation between dissolved solids concentration and "residue on evaporation" or ROE, an alternate measure of the amount of dissolved material in a water sample. Either could be chosen to model as a function of the other, and usually a description of their relationship is what is of most interest.

It is easy to show, however, that the two possible OLS lines (Y on X and X on Y) differ in slope and intercept. Following equation [10.7], reversing the usual order and setting X as the response variable, the resulting OLS equation will be

$$X_i = \bar{X} + r\,\frac{s_x}{s_y}\,(Y_i - \bar{Y}) \qquad [10.8]$$

which when solved for Y becomes

$$Y_i = \bar{Y} + \frac{1}{r}\,\frac{s_y}{s_x}\,(X_i - \bar{X}) \qquad [10.9]$$

Let $b_1' = (1/r \bullet s_y/s_x)$, the slope of X on Y re-expressed to compare with slope b_1. Contrasting [10.7] and [10.9], the slope coefficients $b_1 \neq b_1'$. Thus the two regression lines will differ unless

the correlation coefficient r equals 1.0. In figure 10.7, these two regression lines are plotted for the dissolved solids and ROE data of Appendix C12.

The choice of which, if either, of the OLS lines to use follows a basic guideline. If one is to be predicted from the other, the predicted variable should be assigned as the response variable Y. Errors in this variable are being minimized by OLS. However, when only a single line describing the functional relationship between the two variables is of interest, neither OLS line is the appropriate approach. Neither OLS line uniquely or adequately describes that relationship. A different linear model having a unique solution should be used instead -- the line of organic correlation.

10.2.2 Line of Organic Correlation

The line of organic correlation (LOC) was proposed as a linear fitting procedure in hydrology by Kritskiy and Menkel (1968) and applied to geomorphology by Doornkamp and King (1971). Its theoretical properties were discussed by Kruskal (1953). The line also has been called the "geometric mean functional regression" (Halfon, 1985), the "reduced major axis" (Kermack and Haldane, 1950), the "allometric relation" (Teisser, 1948) and "Maintenance of Variance - Extension" or MOVE (Hirsch, 1982). It possesses three characteristics preferable to OLS in specific situations:

a) LOC minimizes errors in both X and Y directions.
b) It provides a unique line identical regardless of which variable, X or Y, is used as the response variable, and
c) The cumulative distribution function of the predictions, including the variance and probabilities of extreme events such as floods and droughts, estimates those of the actual records they are generated to represent.

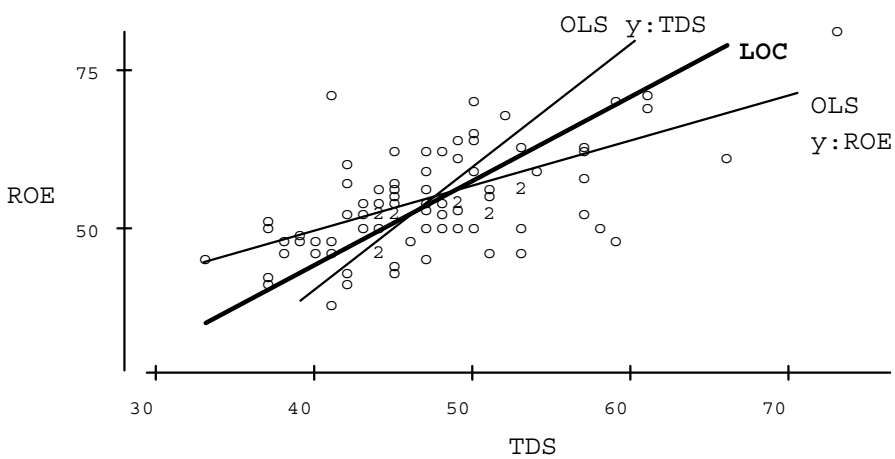


Figure 10.7 Three straight lines fit to the same data.

The LOC minimizes the sum of the areas of right triangles formed by horizontal and vertical lines extending from observations to the fitted line (figure 10.8). By minimizing errors in both directions it lies between the two OLS lines on a plot of Y versus X (see figure 10.7). The slope of the LOC line equals the geometric mean of the Y on X and X on Y OLS slopes:

$$b_1'' = \sqrt{b_1\, b_1'} = \text{sign}[r] \bullet \frac{s_y}{s_x}$$

where b_1'' is the slope of the LOC line

$$Y_i = b_0'' + \text{sign}[r] \bullet \frac{s_y}{s_x} \bullet X_i \qquad [10.10]$$

So the correlation coefficient in the equation for OLS slope is replaced by the algebraic sign (+ or –) of the correlation coefficient with LOC. The magnitude of the LOC slope b_1'' is determined solely by the ratio of standard deviations s_y/s_x. Performing LOC of X on Y will give the identical line as does the LOC of Y on X .

LOC is therefore used for two purposes, corresponding to the three above attributes:

a , b) to model the correct functional relationship between two variables, both of which are measured with error.

c) to produce a series of estimates $\hat{Y}_i$ from observed X_i whose distributional properties are similar to those expected had the Y_i been measured. Such estimates are important when the probability distribution (variance or percentiles) of the estimates, and not just the mean or an individual estimate, are to be interpreted and used.

Examples of the first use for LOC include the geomorphic relationships cited above, describing the relation between bioaccumulation and octanol-water partition coefficients (Halfon, 1985), or other applications where the slope is to take on physical meaning rather than interest in prediction of values of one variable.

One example of the second use for LOC is the extension or fill-in of missing observations. This use for record extension has been the major application of LOC to water resources thus far. As an example, suppose two nearby sites overlap in their gaged record. The streamflow for the site with the shorter record is related to that at the longer (the "base") site during the overlap period. Using this relationship, a series of streamflow data at the shorter site is estimated during an ungaged period based on flows at the base site. If the OLS equation were used to estimate streamflows, the variance of the resulting estimates would be smaller by a factor of R^2 than it should be. OLS reduces the variance of estimates because the OLS slope is a function not only of the ratio of the standard deviations s_y/s_x, but also of the magnitude of the correlation coefficient r. Only when $|r| = 1$ do OLS estimates posses the same variance as would be expected based on the ratio of variances for the original data. To see this more clearly, take the extreme case where r=0, and there is no relationship between

Y and X. The slope then equals 0, and all OLS estimates would be identical and equal to $\overline{Y}$. The variance of the estimates is also zero. As R^2 decreases from 1 to 0, the variance of OLS estimates is proportionately reduced. This variance reduction is eliminated from LOC by eliminating the correlation coefficient from the equation for slope. The estimates resulting from the LOC have a variance in proportion to the ratio of the variances s_y^2/s_x^2 from the original data.

When multiple estimates are to be generated and statements made about probabilities of exceedance, such as flood-flow probabilities, probabilities of low-flows below a water supply intake, or probabilities of exceeding some water-quality standard, inferences are made which depend on the probability distribution of the estimated data. In these cases LOC, rather than OLS, should be used to generate data. OLS estimates would substantially underestimate the variance because they do not include the variability of individual values around the regression line (Hirsch, 1982). As a consequence, the frequency of extreme events such as floods, droughts, or exceedance of standards would be underestimated by OLS.

Variations on using LOC for hydrologic record extension have been published by Vogel and Stedinger (1985) and Grygier et al. (1989).

All three of the lines discussed thus far have two identical characteristics. They are invariant to scale changes, so that changing the Y or X scale (from English to metric units, for example) will not change the estimates of slope or intercept after re-expressing them back into their original scales. However, if the X and Y axes are rotated and lines re-computed, the second set of estimates will differ from the first following re-expression into the original orientation. This second property is not desirable when the original axes are of arbitrary orientation, such as for latitude and longitude. The line discussed in the next section can be fit when invariance to spatial orientation is desired.

10.2.3 Least Normal Squares

Least normal squares is the line which minimizes the squared distances between observed points and the line, where distances are measured perpendicular (normal) to the line. The slope can be expressed as in figure 10.8

$$b = -A + \frac{\sqrt{r^2 + A^2}}{r} \qquad \text{where } A = \frac{1}{2}\left(\frac{s_x}{s_y} - \frac{s_y}{s_x}\right). \qquad [10.11]$$

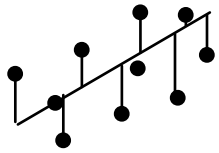
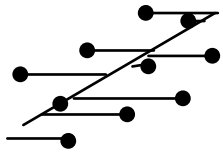
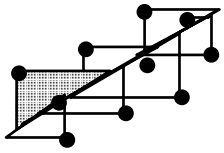
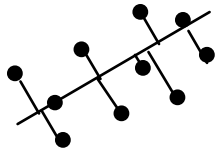
Method	Minimizes:	Slope	Scale Change	Rotation
OLS Y on X		$b_1 = r\frac{s_y}{s_x}$	invariant	changes
OLS X on Y		$b_1' = \frac{1}{r}\frac{s_y}{s_x}$	invariant	changes
LOC		$b_1'' = \text{sign}[r]\ \frac{s_y}{s_x}$	invariant	changes
LNS		$b = -A + \frac{\sqrt{r^2 + A^2}}{r}$ where $A = 0.5\left(\frac{s_x}{s_y} - \frac{s_y}{s_x}\right)$	changes	invariant

Figure 10.8 Characteristics of four parametric methods to fit straight lines to data

An appealing property of LNS is its invariance to rotation of axes. This is desirable when the coordinate system in which the data are measured is arbitrary. The most common example of this is where X and Y are physical locations, such as latitude and longitude. If the axes are rotated, the X and Y coordinates of the data recomputed, and the LNS line recomputed, it will coincide exactly with the LNS line for the data prior to rotation. This is not so with OLS or LOC. However, the LNS line is not invariant to scale changes. The LNS line expressed in any scale will differ depending on the scale in which the calculations were made. For example, the LNS line relating concentration in mg/L to streamflow in cubic feet per second will differ from the LNS line for the same data using streamflow in cubic meters per second. This attribute makes LNS poorly suited to describe the relation between most common water resources variables. Where LNS is appropriate is in computing trajectories minimizing distances between observed points in space. Kirby (1974) thus used LNS to compute the straight line traverse of a ship from a set of coordinate locations taken along its trip.

10.2.4 Summary of the Applicability of OLS, LOC and LNS

To summarize the application of each of the above parametric procedures:

1. To estimate individual values of one variable from another variable, use OLS (assuming the data are linear and homoscedastic). This holds regardless of causality, and regardless of whether there are errors in measurement of the explanatory variable.
2. To estimate multiple values of one variable from another variable in order to make statements about the probability distribution, use LOC. This preserves the characteristics of the entire distribution, avoiding the downward bias in variance of the OLS estimates.
3. To describe the functional relationship between two variables with the primary interest in the slope coefficient, use LOC.
4. To determine the geographic trajectory which minimizes the differences from observed data, use LNS.

10.3 Weighted Least Squares

Data may exhibit a linear pattern yet have non-constant variance (heteroscedasticity -- see figure 10.9). Corrections for non-constant variance involving a power transformation will often alter the linear pattern to one which is curved. Also, transformation into differing units may not be desirable, due to retransformation bias of the estimates (see Chapter 9). Finally, the data may have known inherent differences in their variances, such as when means or other summary statistics based on unequal-sized data sets are used as the explanatory variable. When the constant variance assumption of OLS is violated, an alternate method called weighted least squares (WLS) should instead be employed.

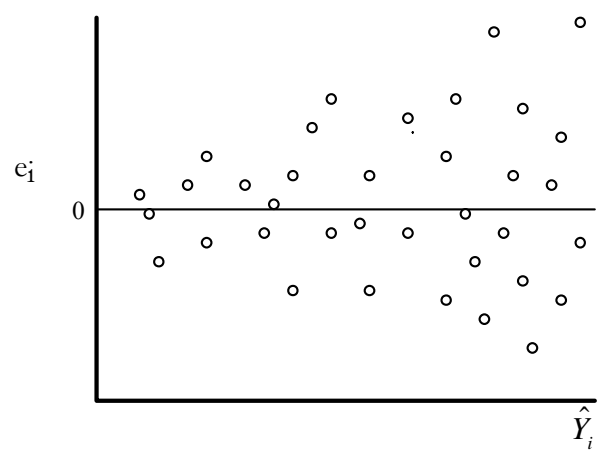


Figure 10.9 Heteroscedastic data.

With WLS, each squared residual $(Y_i - \hat{Y}_i)^2$ is weighted by some weight factor w_i in such a way that observations with greater variance have lesser weight. Thus "less reliable" observations have less influence on the resulting linear equation than "more reliable" observations. The fitted WLS equation minimizes the squares of the <u>weighted</u> residuals. To evaluate whether this weighting has corrected for heteroscedasticity, a weighted residuals plot should be drawn. In this plot the weighted residuals, $e_i\sqrt{w_i}$ are plotted versus $\hat{Y}_i\sqrt{w_i}$. The pattern of weighted residuals can be interpreted as with any other residuals plot.

One common use of WLS in water resources arises when basin characteristics are used to estimate flood percentiles (Tasker, 1980). For example, estimates of the 100-year flood at ungaged sites can be made from a log-log regression of sample estimates of 100-year floods for gaged sites within a region versus drainage area. The flood flows used to construct the regression will have differing variances for different sites, depending on their record lengths n. Sample estimates based on longer records are more reliable, and will have lower variance, than for stations with less data. Therefore estimates from longer records should be given a stronger effect on the regression line. If all original observations are assumed to have constant variance σ^2, then the weights w_i for the weighted regression will be proportional to the record lengths n_i at each station.

Further weighting could reflect any spatial correlation between the sites. This is called generalized least squares, and is applied to hydrology by Stedinger and Tasker (1985). An example of weighting in response to differential sampling within a stratified sampling design is given by DuMouchel and Duncan (1983).

A more empirical method of weighting occurs by setting weights inversely proportional to the sample variance of the response variable at that location. This variance is rarely known ahead of time, so that weights are computed based on residuals from an ordinary least squares regression (OLS) in the following manner:

1) OLS regression is computed for Y versus X. Residuals are plotted against $\hat{Y}$, and nonconstant variance is seen.
2) Observations with similar X's are grouped, and the variance of the observations in each group s_y^2 is calculated. These variances are plotted versus X_i for each group.
3) Assign s_y^2 to each observation in group i. Weights $w_i = 1/s_y^2$.

Weighted least squares can be computed using software for unweighted multiple regression by employing a data transformation $Y_i' = c_i Y_i$, where each observation Y_i is multiplied by the square root of the weight for that point ($c_i = \sqrt{w_i} = 1/s_y$). The X_i must also be weighted as $X_i' = c_i X_i$. A weighted intercept term must also be included as a new "variable" I_i', consisting of a vector of c_i's, one per observation. The transformed Y_i' are then related by multiple regression to X_i' and I_i' using the "no intercept" option (the I' column is the weighted intercept). The resulting coefficients are the coefficients of the weighted least squares line.

<u>Example 3</u>

Total dissolved solids (TDS) from Appendix C12 are plotted versus time, and an increasing variance is seen (figure 10.10). Regression of TDS versus time produces:

$$TDS = -1627 + 0.844 \bullet Time, \quad \text{t-statistic} = 4.62 \qquad p = <0.001$$

where Time is in years. A residuals plot would also show increasing variance.

However, this equation puts undue emphasis on the more recent data, which have the largest variability. The variability seems to increase after 1985, therefore the data are split into two periods, and the variance of TDS is computed separately for each period.

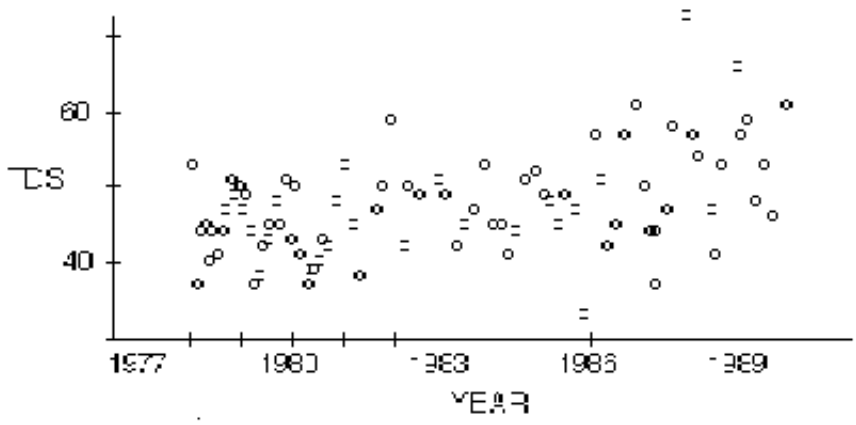


Figure 10.10 TDS data with non-constant variance (heteroscedasticity).

The variance for the pre-1985 period is 24.18, while after 1985 it is 71.80. The reciprocal of these values is assigned as the weight function for each observation in the respective groups, and a weighted least squares regression is performed. This results in:

TDS = −1496 + 0.778 • Time. t-statistic = 4.10 p = <0.001

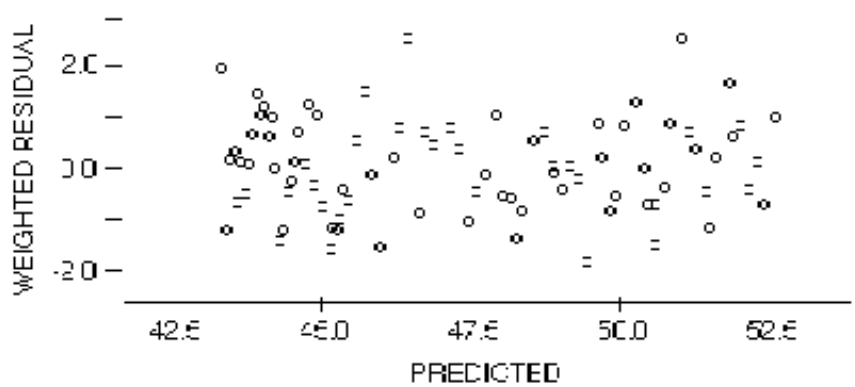


Figure 10.11 Weighted residuals plot of the TDS data.

A plot of the weighted residuals versus predicted values is shown in figure 10.11. The weighted residuals have constant variance. Thus the weighted least squares line should be preferred to the unweighted line, because it more closely conforms to one of the assumptions of least squares regression -- constant variance of residuals.

10.4 Iteratively Weighted Least Squares

OLS regression can be thought of as a "linear mean", with both desirable and undesirable properties similar to a mean. One undesirable property is that outliers can "pull" the location of the line (estimates of slope and intercept) in their direction, much in the same fashion as the sample mean is affected by an outlier. The resulting residuals corresponding to the outlying point may be small, making that point difficult to discern as unusual. Such outliers must be detected using influence statistics (see Chapter 9). In addition to detecting outliers, it may be desirable to limit their influence on the regression line, similar in objective to the Kendall-Theil method given in section 10.1. A second method for doing so, somewhat analogous to a trimmed mean, is a robust regression method called iteratively weighted least squares (IWLS). Unlike Kendall-Theil, IWLS is applicable in the multiple regression context.

The goal of any robust regression is to fit a line not strongly influenced by outliers. This leaves large residuals for the outliers, but a better fit to most other points. IWLS produces models similar to OLS when the underlying residuals distribution is normal, where OLS would have been reasonable to use. Alternate methods of robust regression to IWLS include "least median of squares" and "least absolute value" (Rousseeuw and Leroy, 1987), both of which minimize a more robust measure of error than least squares.

With IWLS, weights are derived from the data. An OLS is first computed -- all weights are initially set equal to one. Points nearest the OLS line are then given weights near one, while points further away have lesser weight. A weighted least squares is computed, and the process repeated. After about two iterations the weights become stabilized, and the final iteratively weighted least squares line results.

There are several weight functions which have been used to compute weights. A common and useful one is the bisquare weight function (Mosteller and Tukey, 1977):

$$w_i = \begin{cases} (1 - u_i^2)^2 & \text{for } |u_i| \le 1 \\ 0 & \text{for } |u_i| > 1 \end{cases}$$

where $u_i = \dfrac{Y_i - \hat{Y}_i}{c \bullet S}$

c = constant, and

S = some robust measure of spread of the residuals $(Y_i - \hat{Y}_i)$.

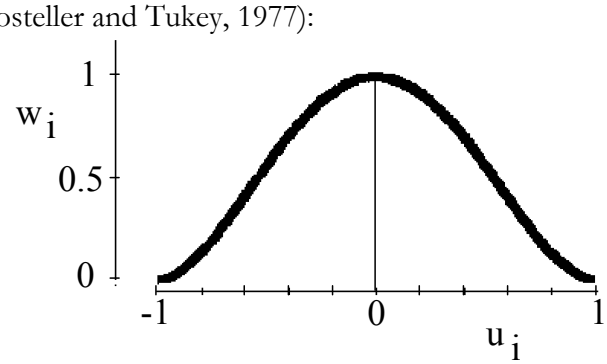


Figure 10.12 Bisquare Weight Function

The purpose of the divisor $c \bullet S$ is to make u_i invariant to scale changes.

Common choices for c and S are

a. c = 3 and S = the IQR of the residuals. For a normal distribution IQR ≅ 4/3 σ, so that when c = 3, c•S ≅ 4σ. This is a margin sufficiently wide to include most or all observations when the distribution is near-normal, and yet protect against outliers when the distribution is markedly non-normal.

b. c = 6 and S = the MAD, the median absolute deviation from the previous line, or median |residuals|. Again c•S ≅ 4σ (MAD ≅ 2/3 σ for a normal distribution).

Note that since the sample standard deviation is strongly distorted by outliers, it would be a poor choice as the measure of spread S. This highlights the failing of all parametric tests for outliers: if the criteria for declaring a value as an outlier is strongly influenced by those same outliers, it will be inflated to the point of declaring too few data as outliers. Either the MAD or IQR are more appropriate than the standard deviation for this purpose.

After calculating the IQR or MAD of residuals from an OLS, the first set of weights are produced. These weights are used in the first weighted least squares, from which new residuals are used to compute new weights. The process is repeated until the weights stabilize -- in most cases only two iterations are required.

<u>Example 3</u>

TCE concentrations were measured in wells from the Upper Glacial Aquifer, Long Island, NY., and related to population density (Eckhardt et al., 1989). Below are listed the percent of wells with TCE concentrations above the detection limit (%DET), by population density of the surrounding land (POPDEN). Compute the robust regression equation (2 iterations) to predict detection percentage from population density.

%DET	0.64	4.80	10.20	22.50	25.00	25.00	67.00	38.00	31.30
POPDEN	1	2	3	5	6	8	9	11	13

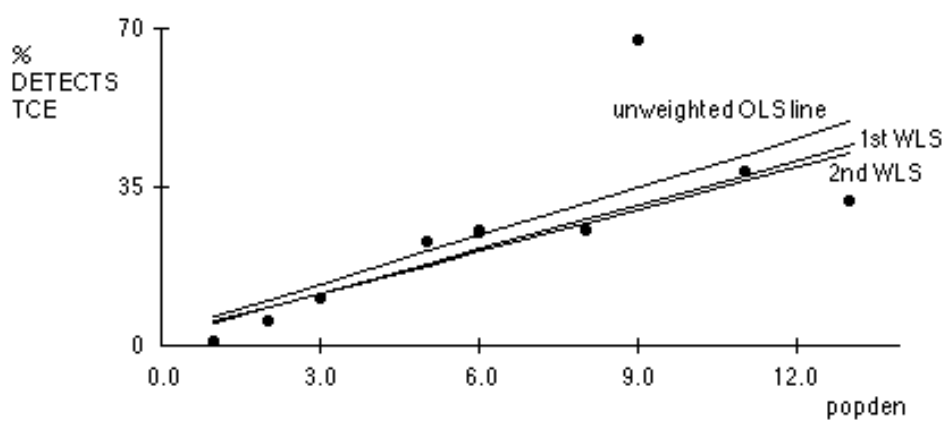


Figure 10.13 TCE concentrations on Long Island (Eckhardt, 1989)

The OLS (unweighted) regression equation is %DET = 2.00 + 3.56 • POPDEN with a t-statistic of 2.86. This line is pulled up by the one outlier at a population density of 9 which doesn't fit the rest of the data very well (figure 10.13). The residuals e_i from this OLS line are used to establish bisquare weights for the first WLS line.

e_i:	−0.414	−0.345	−0.191	0.199	0.120	−0.404	2.478	−0.253	−1.545
w_i:	0.929	0.945	0.982	0.978	0.992	0.913	0.000	0.971	0.326

The outlying point is sufficiently far from the line that it receives a weight of zero. The first weighted regression equation is then %DET = 0.93 + 3.23•POPDEN, with a t-statistic of 6.93. This is shown as "1st WLS" in figure 10.13. Again, residuals are computed from this equation, and a new set of weights computed:

w_i:	0.945	0.970	0.999	0.872	0.903	0.986	0.000	0.989	0.489

The 2nd iteration weighted regression equation is then %DET = 1.24 + 3.10•POPDEN, similar to the previous iteration, with a t-statistic of 6.63. Figure 10.13 shows this line as "2nd WLS". The residual for the outlying point remains large, while the line fits the majority of the data quite well. This is the objective of a robust regression.

10.5 Smoothing

Smoothing differs in purpose and form from the previous methods. It is an exploratory technique, having no simple equation or significance tests associated with it. The most common smooths estimate the center of the data -- the conditional mean or median of Y as X changes. The lack of an equation is a strength in the sense that a smooth is not constrained by some prior assumption as to the mathematical function of the relationship. Rarely are there theoretical grounds for choosing one function over another in modeling Y versus X. For large data sets it is common to visually identify departures from a simple function which could only be modelled by incorporating several high order terms. This can cause instability near or beyond the range of the data. The shape of a smooth is not specified *a priori*, but is determined solely by the data.

Middle smooths allow the data to dictate the location of a smooth curve which goes through the middle of the data. They are used to highlight trends or patterns in the data on a scatterplot. These patterns are often difficult to see. The human eye only poorly follows the central tendency of a scatterplot; the range of data dominates visual impression. Adding a line through the middle draws attention to the center of the plot, aiding judgement of whether the pattern is linear, indicating where breaks in slope occur, etc.

10.5.1 Moving Median Smooths

The simplest smooths are moving averages or medians. Data are smoothed by calculating the mean or median of a portion of the total data within some 'window' of influence around a given

X_0. This is repeated while setting X_0 equal to nearly every X value in the data set. As before, outliers will influence moving averages (means) more strongly than medians, so that moving averages are more erratic than medians in the vicinity of outliers. Moving medians therefore are more resistant to outliers than are moving averages.

Suppose a 5-point moving median is to be computed. A 'window' of width equal to 5 data points is begun at the left of the X-Y plot. The median of the 5 Y values within the window is computed, and plotted at the center of the window (X_0 = 3rd point from the left) to form the first value of the smooth. Data outside the window have no influence on the smoothed value. The X window is shifted to the right by one data point, a new median of the 2nd through 6th points calculated, and this value plotted at the new X_0 = 4th point from the left. This shifting and computation progressively continues through the final window, composed of the rightmost 5 points. All medians are then connected by straight lines to form the moving-median smooth.

Figure 10.14 shows an 11-point moving median smooth for sand concentrations in the Colorado River at Lees Ferry, Arizona. Moving medians are convenient for hand computation, but produce a "rough" pattern unless the window size is quite large. Large windows result in the undesirable characteristic that data far from X_0 influence the resulting value as much as data nearby. To avoid this, more complex smoothing routines are now performed by computer.

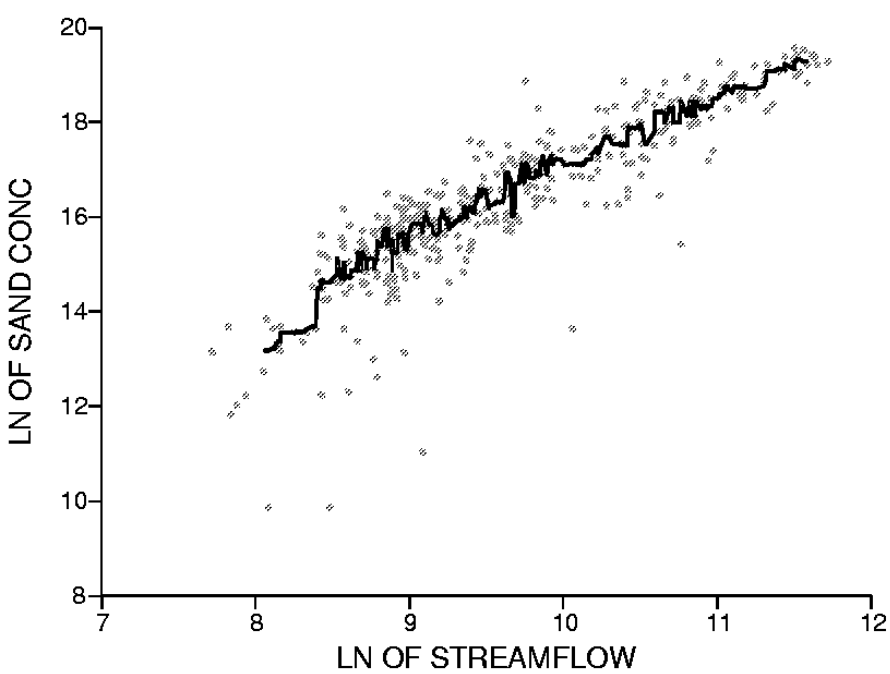


Figure 10.14 11 point moving median of sand concentrations in the Colorado River at Lees Ferry, Arizona, 1949-1970.

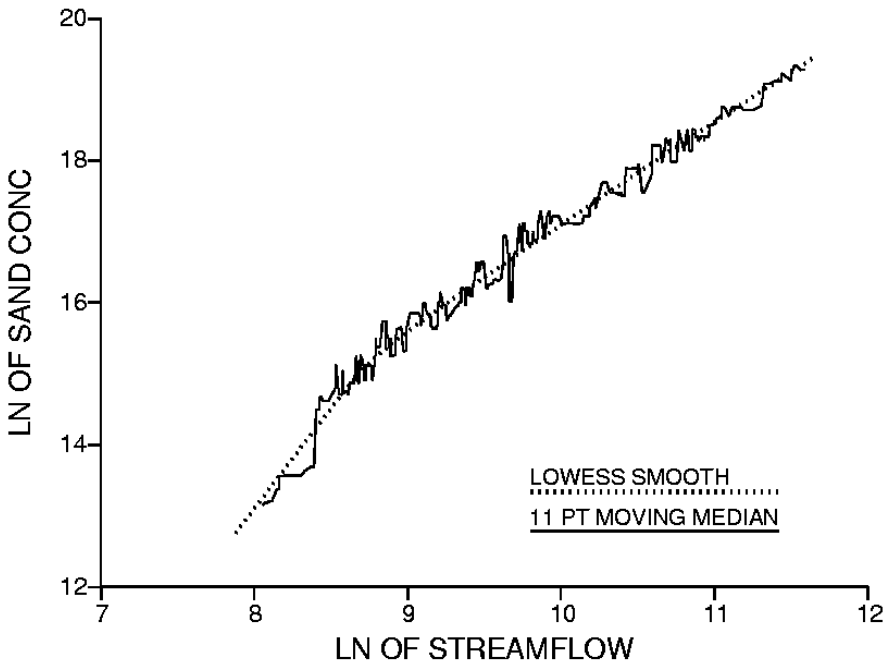


Figure 10.15 11 point moving median and LOWESS smooths of the Lees Ferry data

These allow the data nearer the center of the window to influence the smoothed value more than those further away. They also allow the smoothness of the final fit to be adjusted to the needs of the data analyst. One of the most flexible and useful smoothing algorithms is called LOWESS. In figure 10.15 the 11 point moving median smooth is compared to a LOWESS smooth for the Lees Ferry data.

10.5.2 LOWESS

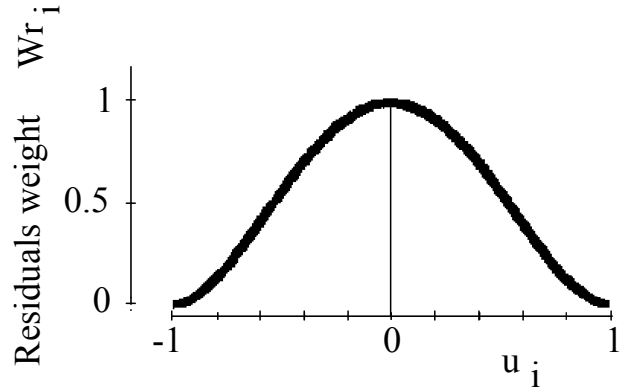
LOWESS, or LOcally WEighted Scatterplot Smoothing (Cleveland et al., 1979) is computationally intensive. It involves fitting at least 2^n weighted least squares equations. At every X_0, a $\hat{Y}$ is computed from a WLS regression whose weights are a function of both the distance from X_0 and the magnitude of the residual from the previous regression (an iterative procedure). The robust regression weights w_i are computed by

$$w_i = wx_i \bullet wr_i$$

where wx_i, the distance weight, is a function of the distance between the center of the window X_i and all other X. The residuals weight wr_i is a function of $|Y_i - \hat{Y}_i|$, the distance in the Y direction between the observed Y_i and the value predicted from the previous WLS equation. A point will receive a small weight, and therefore have little influence on the smoothed $\hat{Y}$, if it is either far from the center of the window in the X direction or has a large residual in the Y direction. The measure of how quickly weights decrease as distances increase in the X and Y directions is determined by the weight function. For a point at (X_i,Y_i), the bisquare weight is determined as

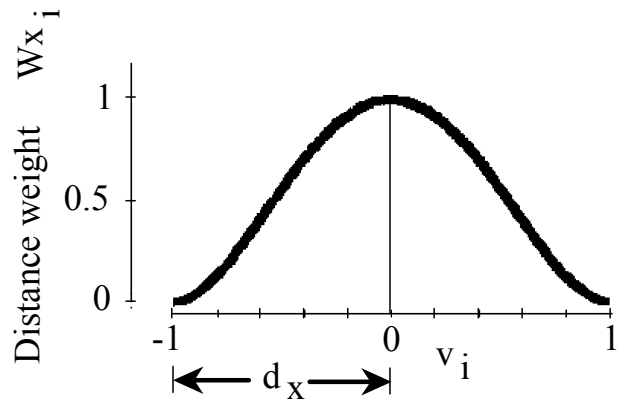
$$wr_i = \begin{cases} (1 - u_i^2)^2 & \text{for } |u_i| \le 1 \\ 0 & \text{for } |u_i| > 1 \end{cases}$$

$$\text{where } u_i = \frac{Y_i - \hat{Y}_i}{6 \bullet \text{median of all } |Y_i - \hat{Y}_i|}$$



$$wx_i = \begin{cases} (1 - v_i^2)^2 & \text{for } |v_i| \le 1 \\ 0 & \text{for } |v_i| > 1 \end{cases}$$

$$\text{where} \quad v_i = \frac{X_i - X}{d_x}$$



where d_x = half width of window = mth largest $|X_i - X|$

$m = Nf$

N = sample size

f = smoothness factor specified at outset.

Smoothness of LOWESS is varied by altering the window width, as controlled by the smoothness factor f (figure 10.16). As f is increased, the window size is increased, and more points influence the magnitude of $\hat{Y}$. Selection of an appropriate f is determined subjectively according to the purpose for which the smooth is used.

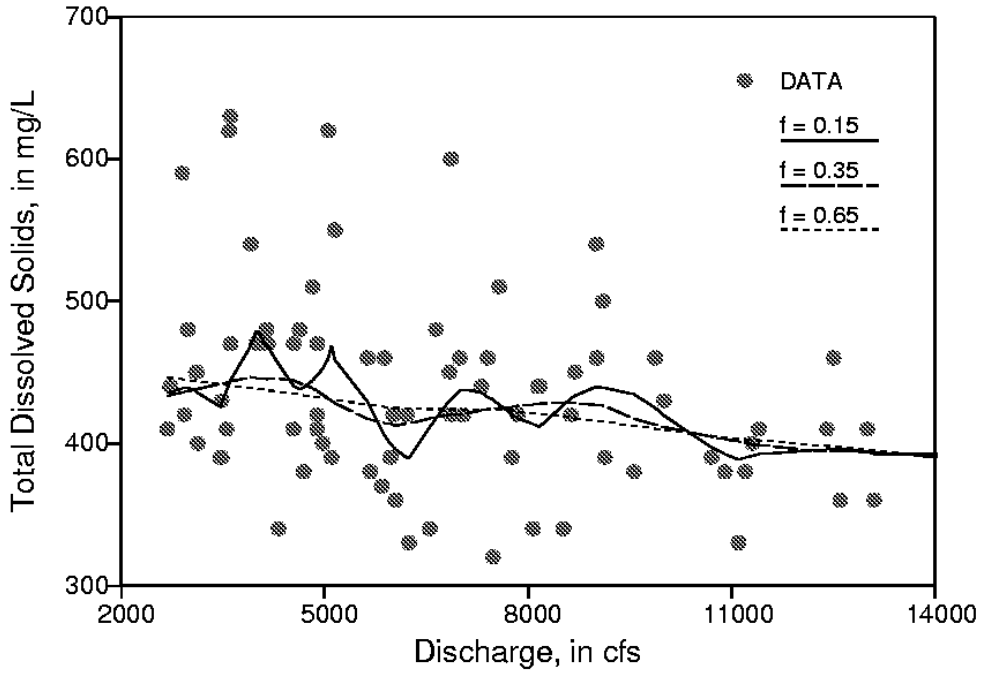


Figure 10.16 Three smooths of the same data with differing smoothness factors f.

Three examples of situations in which LOWESS smooths greatly aid data analysis are:

1. To emphasize the shape of the relationship between two variables on a scatterplot of moderate to large sample size. Adding a line through the middle draws attention to the center of the plot, aiding judgement of how the two variables are related.

2. To compare and contrast multiple large data sets. Plotting all data points with differing symbols per group does not provide the clarity necessary to distinguish similarities and differences between groups. Instead, computing and plotting LOWESS smooths without the data may give great insight into group characteristics. For example, Welch et al. (1988) used LOWESS to describe the relationship between arsenic and pH in four physiographic regions of the Western United States (figure 2.26 in Chapter 2). Thousands of data points were involved; a scatterplot would have shown nothing but a blob of data. The smooths clearly illustrated that in three regions arsenic concentrations increased with increasing pH, while in the fourth no increase was observed. Smooths were also used by Schertz and Hirsch (1985) to illustrate regional patterns in atmospheric precipitation chemistry. They used one smooth per station to display simultaneous changes in sulfate and other chemical concentrations occuring over broad regions of the country (figure 10.17). These relationships would have gone unnoticed using scatterplots -- the underlying patterns would have been obscured by the proliferation and scatter of the data.

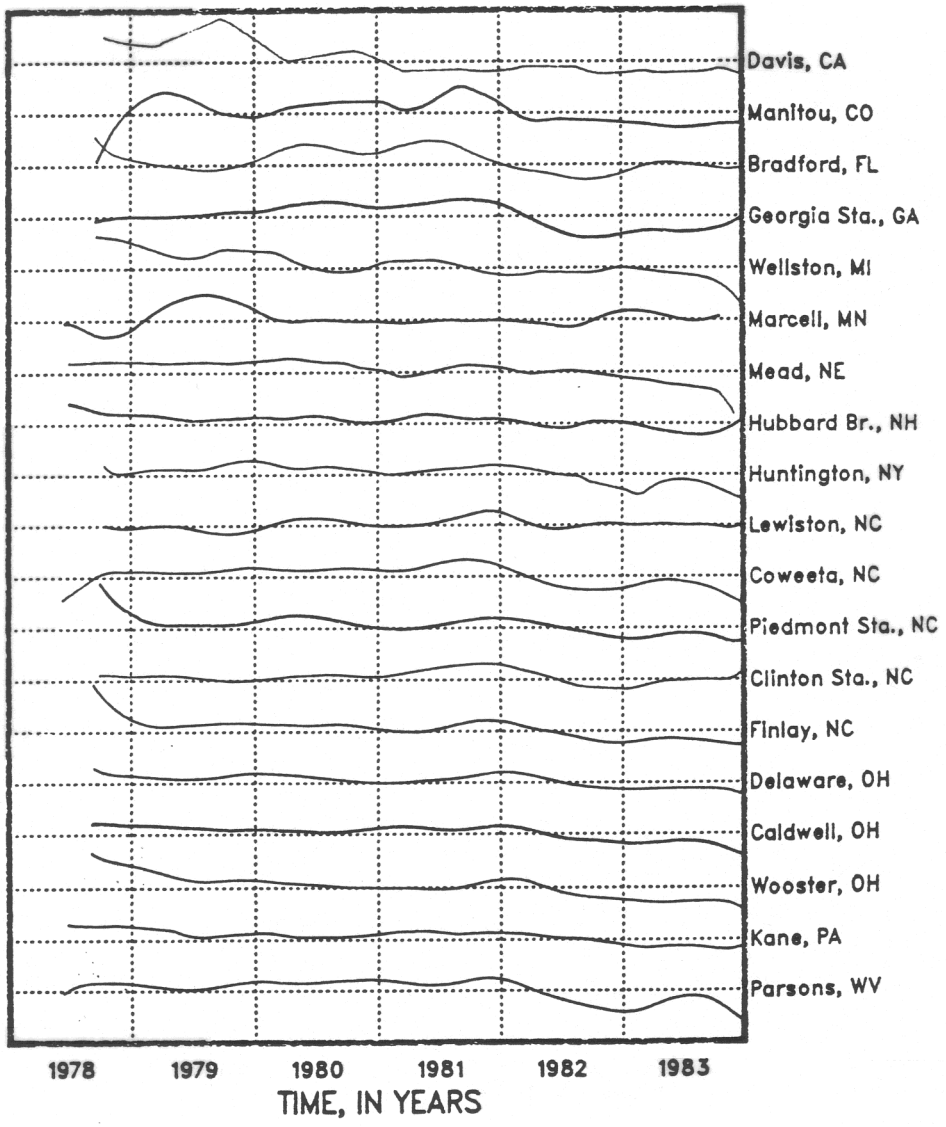


Figure 10.17 Smooths of sulfate concentrations at 19 stations, 1978-83 (from Schertz and Hirsch, 1985).

3. To remove the effect of an explanatory variable without first assuming the form of the relation (linear, etc.). In situations equivalent to multiple regression where several variables may affect the magnitude of a response variable (Y), removal of one variable's (X) effects may be accomplished by computing a LOWESS smooth of Y versus X and using the residuals from the smooth in subsequent analyses. An example is when removing the effects of discharge or precipitation volume from chemical concentration data prior to performing a trend analysis (see Chapter 12). LOWESS allows the analyst to be unconcerned as to whether the relation between Y and X is linear or nonlinear. In contrast, linearity would have to be established prior to using regression.

Two additonal lines are sometimes plotted along with the LOWESS middle smooth. These are upper and lower smooths (Cleveland and McGill, 1984b), which function as smoothed versions of upper and lower quartiles of the conditional distribution of Y as a function of X. They are constructed by computing additional LOWESS smooths on the positive residuals and negative residuals, respectively, from the middle LOWESS smooth. These values are then added to the middle smooth, and connected with straight line segments. Upper and lower smooths are useful for showing how the spread and/or symmetry of the conditional distribution of Y changes as a function of X. Figure 10.18 is one example. It shows how the spread of nitrate concentrations changes with depth for groundwaters under Long Island, NY. The spread or "running IQR" is indicated by the distance between the upper and lower smooths, shown as dashed lines in the plot.

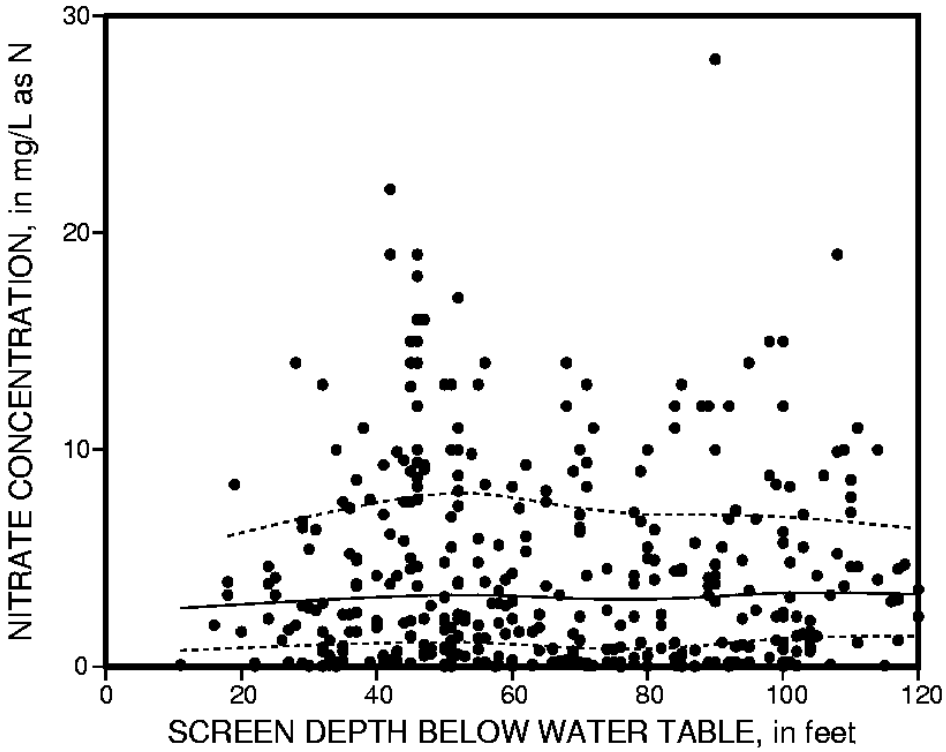


Figure 10.18 Nitrate concentrations versus depth in the upper Glacial Aquifer, Long Island NY (data from Eckhardt et al., 1989).

10.5.3 Polar Smoothing

Polar smooths (Cleveland and McGill, 1984b) are variations on lowess smooths. They are polygons describing the two-dimensional locations of data groups on a scatterplot (see figure 2.28 in Chapter 2). Comparisons of differences in location of several data groups is made much easier by comparing polar smooths rather than comparing symbols for each data point on a scatterplot, as in figure 2.27. Polar smooths are used as a visual 'discriminant analysis' in two dimensions.

To compute a polar smooth, first center the data at the median of X and median of Y. All data points are then described in terms of their angle and radius from this center, placing the data into polar coordinates. A lowess smooth is computed while in polar coordinates, and then is re-transformed back into original units. The smooth, which while in polar coordinates had 50 percent of the data below it, upon re-transformation envelops those same 50 percent within it An analogous 'upper smooth' which in polar coordinates had 75 percent of the data below it becomes an 'outer smooth' containing 75 percent of the data in original units.

Polar smooths can be a great aid to exploratory data analysis. They are not constrained a priori to be an ellipse or any other shape, but take on the characteristics of the data. This can lead to new insights difficult to see by plotting the original observations. For example, in figure 2.28 smooths enclosing 75% of the conductance versus pH data for three types of upstream land use are plotted. The irregular pattern for the smooth of abandoned mine data suggests that two separate subgroups are present, one with higher pH than the other.

Exercises

10.1 For the data below,

a) compute the Kendall slope estimator,

b) compute Kendall's τ,

c) compute the non-parametric regression equation.

d) compute the significance level of the test.

Y	10	40	30	55	62	56
X	1	2	3	4	5	6

10.2 One value has been altered from the 10.1 exercise. Again compute the slope estimate, intercept, τ and significance level. By how much have these changed in response to the one (large) change in Y? Also compute a 95% confidence interval on the slope estimate.

Y	10	40	30	55	200	56
X	1	2	3	4	5	6

10.3 Compute the robust IWLS equation (2 iterations) for the Exercise 10.2 data.

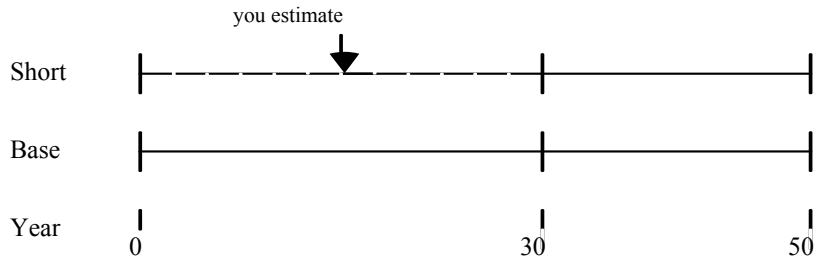
10.4 Williams and Wolman (1984) relate the lowering of streambed elevation downstream of a major dam to years following its installation. Calculate a linear least-squares regression of bed lowering (L) as the response variable, versus years (Yrs) as the explanatory variable, and compute its R^2.

Yrs	Lowering (m)	Yrs	L	Yrs	L
0.5	−0.65	8	−4.85	17	−5.05
1	−1.20	10	−4.40	20	−5.10
2	−2.20	11	−4.95	22	−5.65
4	−2.60	13	−5.10	24	−5.50
6	−3.40	15	−4.90	27	−5.65

Calculate a 5-point moving median smooth of the data. Plot the smooth and regression line along with a scatterplot of the data. Describe how well each represents the data.

10.5 Record Extension

Monthly discharges for September at two rivers are given in Appendix C13 (units of million cubic meters per month). The most recent 20 years are available for "Short" (ignore the data in italics), and 50 years at "Base". The two sites are close enough that the data are reasonably well correlated with each other. Using the 20 years of joint record and the additional 30 years of record at "Base", produce a 50-year-long record at "Short" for use in a water supply simulation model.



First use regression and then repeat the process using the LOC. Take the extended record (the 30-year estimates plus the known 20 years) produced by the two methods at "Short" and plot them to illustrate the differences (a boxplot or probability plot are recommended). Compare these to each other and to a plot of the flows which actually occurred (the true flows are given in italics in Appendix C13). Which technique is preferable if the objective is to estimate water supply shortage risks? Which technique is preferable if the objective is to estimate the true September flow in each year? Quantify your conclusion about this.

10.6 The pulp liquor waste contamination of shallow groundwater (see Exercise 7.1) is revisited. Now the relationship between pH and COD in samples taken from the piezometers is of interest. Calculate a straight line which best describes the relationship between these two chemical constituents. Should this line be used by the field technician to predict COD from the pH measured on-site?

pH	COD	pH	COD	pH	COD
7.0	51	6.3	21	8.4	283
7.2	60	6.9	17	7.6	2170
7.5	51	7.0	34	7.5	6580
7.7	3600	6.4	43	7.4	3340
8.7	6900	6.8	34	9.3	7080
7.8	7700	6.7	43	9.0	10800

Chapter 11

Multiple Linear Regression

The 100-year flood is to be estimated for locations without streamflow gages using basin characteristics at those locations. A regression equation is first developed relating the 100-year flood to several basin characteristics at sites which have a streamflow gage. Each characteristic used is known to influence the magnitude of the 100-year flood, has already been used in adjoining states, and so will be included in the equation regardless of whether it is significant for any individual data set. Values for the basin characteristics at each ungaged site are then input to the multiple regression equation to produce the 100-year flood estimate for that site.

Residuals from a simple linear regression of concentration versus streamflow show a consistent pattern of seasonal variation. To make better predictions of concentration from streamflow, additional explanatory variables are added to the regression equation, modeling the pattern seen in the data.

As an exploratory tool in understanding possible causative factors of groundwater contamination, data on numerous potential explanatory variables are collected. Each variable is plausible as an influence on nitrate concentrations in the shallowest aquifer. Stepwise or similar procedures are performed to select the "most important" variables, and the subsequent regression equation is then used to predict concentrations The analyst does not realize that this regression model is calibrated, but not verified.

Multiple linear regression (MLR) is the extension of simple linear regression (SLR) to the case of multiple explanatory variables. The goal of this relationship is to explain as much as possible of the variation observed in the response (y) variable, leaving as little variation as possible to unexplained "noise". In this chapter methods for developing a good multiple regression model are explained, as are the common pitfalls such as multi-collinearity and relying on R^2. The mathematics of multiple regression, best handled by matrix notation, will not be extensively covered here. See Draper and Smith (1981) or Montgomery and Peck (1982) for this.

11.1 Why Use MLR?

When are multiple explanatory variables required? The most common situation is when scientific knowledge and experience tells us they are likely to be useful. For example, average runoff from a variety of mountainous basins is likely to be a function both of average rainfall and of altitude; average dissolved solids yields are likely to be a function of average rainfall, percent of basin in certain rock types, and perhaps basin population. Concentrations of contaminants in shallow groundwater are likely to be functions of both source terms (application rates of fertilizers or pesticides) and subsurface conditions (soil permeability, depth to groundwater, etc.).

The use of MLR might also be indicated by the residuals from a simple linear SLR. Residuals may indicate there is a temporal trend (suggesting time as an additional explanatory variable), a spatial trend (suggesting spatial coordinates as explanatory variables), or seasonality (suggesting variables which indicate which season the data point was collected in). Analysis of a residuals plot may also show that patterns of residuals occur as a function of some categorical grouping representing a special condition such as: on the rising limb of a hydrograph, at cultivating time, during or after frontal storms, in wells with PVC casing, measurements taken before 10:00 a.m., etc. These special cases will only be revealed by plotting residuals versus a variety of variables -- in a scatterplot if the variable is continuous, in grouped boxplots if the variable is categorical. Seeing these relationships should lead to definition of an appropriate explanatory variable and its inclusion in the model if it significantly improves the fit.

11.2 MLR Model

The MLR model will be denoted:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

where y is the response variable
β_0 is the intercept
β_1 is the slope coefficient for the first explanatory variable
β_2 is the slope coefficient for the second explanatory variable
β_k is the slope coefficient for the kth explanatory variable, and
ε is the remaining unexplained noise in the data (the error).

To simplify notation the subscript i, referring to the i=1,2,..,n observations, has been omitted from the above. There are k explanatory variables, some of which may be related or correlated to each other (such as the previous 5-day's rainfall and the the previous 1-day rainfall). It is therefore best to avoid calling these "independent" variables. They may or may not be independent of each other. Calling them explanatory variables describes their purpose: to explain the variation in the response variable.

11.3 Hypothesis Tests for Multiple Regression

11.3.1 Nested F Tests

The single most important hypothesis test for MLR is the F test for comparing any two nested models. Let model "s" be the "simpler" MLR model

$$y_s = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon_s .$$

It has k+1 parameters including the intercept, with degrees of freedom (df_s) of n−(k+1). Again, the degrees of freedom equals the number of observation minus the number of parameters estimated, as in SLR. Its sum of squared errors is SSE_s.

Let model "c" be the more complex regression model

$$y_c = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \beta_{k+1} x_{k+1} + ... + \beta_m x_m + \varepsilon_c .$$

It has m+1 parameters and residual degrees of freedom (df_c) of n−(m+1). Its sum of squared errors is SSE_c.

The test of interest is whether the more complex model provides a sufficiently better explanation of the variation in y than does the simpler model. In other words, do the extra explanatory variables x_{k+1} to x_m add any new explanatory power to the equation? The models are "nested" because all of the k explanatory variables in the simpler model are also present in the complex model, and thus the simpler model is nested within the more complex model. The null hypothesis is

H_0: $\beta_{k+1} = \beta_{k+2} = ... = \beta_m = 0$ versus the alternative

H_1: at least one of these m−k coefficients is not equal to zero.

If the slope coefficients for the additional explanatory variables are all not significantly different from zero, the variables are not adding any explanatory power in comparison to the cost of adding them to the model. This cost is measured by the loss in the degrees of freedom = m−k, the number of additional variables in the more complex equation.

The test statistic is

$$F = \frac{(SSE_s - SSE_c) / (df_s - df_c)}{(SSE_c / df_c)} \qquad \text{where } (df_s - df_c) = m-k.$$

If F exceeds the tabulated value of the F distribution with ($df_s - df_c$) and df_c degrees of freedom for the selected α (say α=0.05), then H_0 is rejected. Rejection indicates that the more complex model should be chosen in preference to the simpler model. If F is small, the additional variables are adding little to the model, and the simpler model would be chosen over the more complex.

Note that rejection of H_0 does not mean that all of the K+1 to m variables have coefficients significantly different from zero. It merely states that some of the coefficients in the more complex model are significant, making that model better than the simpler model tested. Other simpler models having different subsets of variables may need to be compared to the more complex model before choosing it as the "best".

11.3.2 Overall F Test

There are two special cases of the nested F test. The first is of limited use, and is called the overall F test. In this case, the simpler model is

$$y_s = \beta_0 + \varepsilon_s \text{ , where } \beta_0 = \bar{y} \text{ .}$$

The rules for a nested F test still apply: the $df_s = n-1$ and SSE_s equals (n−1) times the sample variance of y. Many computer packages give the results of this F-test. It is not very useful because it tests only whether the complex regression equation is better than no regression at all. Of much greater interest is which of several regression models is best.

11.3.3 Partial F Tests

The second special case of nested F tests is the partial F test, which is called a Type III test by SAS. Here the complex model has only 1 additional explanatory variable over the simpler model, so that m=k+1. The partial F test evaluates whether the mth variable adds any new explanatory power to the equation, and so ought to be in the regression model, given that all the other variables are already present. Note that the F statistics on a coefficient will change depending on what other variables are in the model. Thus the simple question "does variable m belong in the model?" cannot be answered. What can be answered is whether m belongs in the model in the presence of the other variables.

With only one additional explanatory variable, the partial F test is identical in results to a t-test on the coefficient for that variable. In fact, $t^2 = F$, where both are the statistics computed for the same coefficient for the partial test. Some computer packages report the F statistic, and some the t-test, but the p-values for the two tests are identical. The partial t-test can be easily performed by comparing the t statistic for the slope coefficient to a students t-distribution with n−(m+1) degrees of freedom. H_0 is rejected if $|t| > t_{1-(\alpha/2)}$. For a two-sided test with $\alpha = 0.05$ and sample sizes n of 20 or more, the critical value of t is $|t| \cong 2$. Larger t-statistics (in absolute value) for a slope coefficient indicate significance. Squaring this, the critical partial F value is near 4.

Partial tests guide the evaluation of which variables to include in a regression model, but are not sufficient for every decision. If every $|t|>2$ for each coefficient, then it is clear that every explanatory variable is accounting for a significant amount of variation, and all should be present. When one or more of the coefficients has a $|t|<2$, however, some of the variables should be removed from the equation, but **the t values are not a certain guide as to which**

ones to remove. These partial t or F tests are precisely the tests used to make automatic decisions for removal or inclusion in "stepwise" procedures: forward, backward, and stepwise multiple regression. These procedures do not guarantee that some "best" model is obtained, as discussed later. Better procedures are available for doing so.

11.4 Confidence Intervals

Confidence intervals can be computed for the regression slope coefficients β_k, and for the mean response $\hat{y}$ at a given value for all explanatory variables. Prediction intervals can be similarly computed around an individual estimate of y. These are entirely analogous to the SLR situation, but require matrix manipulations for computation. A brief discussion of them follows. More complete treatment can be found in many statistics textbooks, such as Montgomery and Peck (1982), Draper and Smith (1981), and Walpole and Myers (1985), among others.

11.4.1 Variance-Covariance Matrix

In MLR, the values of the k explanatory variables for each of the n observations, along with a vector of 1s for the intercept term, can be combined into a matrix X:

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & . & . & x_{1k} \\ 1 & x_{12} & x_{22} & . & . & x_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{nk} \end{bmatrix}$$

X is used in MLR to compute the variance-covariance matrix $\sigma^2 \bullet (X'X)^{-1}$,
where $(X'X)^{-1}$ is the "X prime X inverse" matrix. Elements of $(X'X)^{-1}$ for three explanatory variables are as follows:

$$(X'X)^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} & C_{03} \\ C_{10} & C_{11} & C_{12} & C_{13} \\ C_{20} & C_{21} & C_{22} & C_{23} \\ C_{30} & C_{31} & C_{32} & C_{33} \end{bmatrix} \qquad [11.1]$$

When multiplied by the error variance σ^2 (estimated by the variance of the residuals, s^2), the diagonal elements of the matrix C_{00} through C_{33} become the variances of the regression coefficients, while the off-diagonal elements become the covariances between the coefficients. Both $(X'X)^{-1}$ and s^2 can be output from MLR software.

11.4.2 Confidence Intervals for Slope Coefficients

Interval estimates for the regression coefficients β_0 through β_k are often printed by MLR software. If not, the statistics necessary to compute them are. As with SLR it must be assumed

that the residuals are normally distributed with variance σ^2. A $100\bullet(1-\alpha)\%$ confidence interval on β_j is

$$\hat{b}_j - t_{(\alpha/2,\, n-p)} \sqrt{s^2 C_{jj}} \;\le\; \beta_j \;\le\; \hat{b}_j + t_{(\alpha/2,\, n-p)} \sqrt{s^2 C_{jj}} \qquad [11.2]$$

where C_{jj} is the diagonal element of $(X'X)^{-1}$ corresponding to the jth explanatory variable. Often printed is the standard error of the regression coefficient:

$$\text{se}(\hat{b}_j) = \sqrt{s^2 C_{jj}}\,. \qquad [11.3]$$

Note that C_{jj} is a function of the other explanatory variables as well as the jth. Therefore the interval estimate, like $\hat{b}_j$ and its partial test, will change as explanatory variables are added to or deleted from the model.

11.4.3 Confidence Intervals for the Mean Response

A $100\bullet(1-\alpha)\%$ confidence interval for the expected mean response $\mu(y_0)$ for a given point in multidimensional space x_0 is symmetric around the regression estimate $\hat{y}_0$. These intervals also require the assumption of normality of residuals.

$$\hat{y}_0 - t_{(\alpha/2,n-p)}\sqrt{s^2\, x_0'(X'X)^{-1} x_0} \;\le \mu(y_0) \le\; \hat{y}_0 + t_{(\alpha/2,n-p)}\sqrt{s^2\, x_0'(X'X)^{-1} x_0} \qquad [11.4]$$

The variance of the mean is the term under the square root sign. It changes with x_0, increasing as x_0 moves away from the multidimensional center of the data. In fact, the term $x_0'(X'X)^{-1}x_0$ is the leverage statistic h_i, expressing the distance that x_0 is from the center of the data.

11.4.4 Prediction Intervals for an Individual y

A $100\bullet(1-\alpha)\%$ prediction interval for a single response y_0, given a point in multidimensional space x_0, is symmetric around the regression estimate $\hat{y}_0$. It requires the assumption of normality of residuals.

$$\hat{y}_0 - t_{(\alpha/2,n-p)}\sqrt{s^2\left\langle 1 + x_0'(X'X)^{-1} x_0\right\rangle} \le y_0 \le\; \hat{y}_0 + t_{(\alpha/2,n-p)}\sqrt{s^2\left\langle 1 + x_0'(X'X)^{-1} x_0\right\rangle} \qquad [11.5]$$

11.5 Regression Diagnostics

As was the case with SLR, it is important to use graphical tools to diagnose deficiencies in MLR. The following residuals plots are very important: normal probability plots of residuals, residuals versus predicted (to identify curvature or heteroscedasticity), residuals versus time sequence or location (to identify trends), and residuals versus any candidate explanatory variables not in the model (to identify variables, or appropriate transformations of them, which may be used to improve the model fit).

11.5.1 Partial Residual Plots

As with SLR, curvature in a plot of residuals versus an explanatory variable included in the model indicates that a transformation of that explanatory variable is required. Their relationship should be linear. To see this relation, however, residuals should not be plotted directly against explanatory variables; the other explanatory variables will influence these plots. For example, curvature in the relationship between e and x_1 may show up in the plot of e versus x_2, erroneously indicating that a transformation of x_2 is required. To avoid such effects, partial residuals plots (also called adjusted variable plots) should be constructed.

The partial residual is

$$e_j^* = y - \hat{y}_{(j)}$$

where $\hat{y}_{(j)}$ is the predicted value of y from a regression equation where x_j is left out of the model. All other candidate explanatory variables are present.

This partial residual is then plotted versus an adjusted explanatory variable

$$x_j^* = x - \hat{x}_{(j)}$$

where $\hat{x}_{(j)}$ is the x_j predicted from a regression against all other explanatory variables. So x_j is treated as a response variable in order to compute its adjusted value. The partial plot (e_j^* versus x_j^*) describes the relationship between y and the jth explanatory variable after all effects of the other explanatory variables have been removed. Only the partial plot accurately indicates whether a transformation of x_j is necessary.

11.5.2 Leverage and Influence

The regression diagnostics of Chapter 9 are much more important in MLR than in SLR. It is very difficult when performing multiple regression to recognize points of high leverage or high influence from any set of plots. This is because the explanatory variables are multidimensional. One observation may not be exceptional in terms of each of its explanatory variables taken one at a time, but viewed in combination it can be very exceptional. Numerical diagnostics can accurately detect such anomalies.

The leverage statistic $h_i = x_0'(X'X)^{-1}x_0$ expresses the distance of a given point x_0 from the center of the sample observations (see also section 11.4.3). It has two important uses in MLR. The first is the direct extension of its use in SLR -- to identify points unusual in value of the explanatory variables. Such points warrant further checking as possible errors, or may indicate a poor model (transformation required, relationships not linear, etc.).

The second use of h_i is when making predictions. The leverage value for a prediction should not exceed the largest h_i in the original data set. Otherwise an extrapolation beyond the envelope surrounding the original data is being attempted. The regression model may not fit well in that region. It is sometimes difficult to recognize that a given x_0 for which a predicted $\hat{y}$ is attempted is outside the boundaries of the original data. This is because the point may not be

beyond the bounds of any of its individual explanatory variables. Checking the leverage statistic guards against an extrapolation that is difficult to detect from a plot of the data.

Example 1

Variations in chemical concentrations within a steeply dipping aquifer are to be described by location and depth. The data are concentrations (C) plus three coordinates: distance east (DE), distance north (DN), and well depth (D). Data were generated using $C = 30 + 0.5\,D + \varepsilon$. Any acceptable regression model should closely reproduce this true model, and should find C to be independent of DE and DN. Three pairwise plots of explanatory variables (figure 11.1) do not reveal any "outliers" in the data set. Yet compared to the critical leverage statistic $h_i=3p/n=0.6$, and critical influence statistic $DFFITS=2\sqrt{p/n}=0.9$, the 16th observation is found to be a point of high leverage and high influence (table 11.1). In figure 11.2 the axes have been rotated, showing observation 16 to be lying outside the plane of occurrence of the rest of the data, even though its individual values for the three explanatory variables are not unusual.

Obs. #	DE	DN	D	C	hi	DFFITS
1	1	1	4.2122	30.9812	0.289433	-0.30866
2	2	1	8.0671	33.1540	0.160670	-0.01365
3	3	1	10.7503	37.1772	0.164776	0.63801
4	4	1	11.9187	35.3864	0.241083	-0.04715
5	1	2	11.2197	35.9388	0.170226	0.42264
6	2	2	12.3710	31.9702	0.086198	-0.51043
7	3	2	12.9976	34.9144	0.087354	-0.19810
8	4	2	15.0709	36.5436	0.165040	-0.19591
9	1	3	12.9886	38.3574	0.147528	0.53418
10	2	3	18.3469	39.8291	0.117550	0.45879
11	3	3	20.0328	40.0678	0.121758	0.28961
12	4	3	20.5083	37.4143	0.163195	-0.47616
13	1	4	17.6537	35.3238	0.165025	-0.59508
14	2	4	17.5484	34.7647	0.105025	-0.77690
15	3	4	23.7468	40.7207	0.151517	0.06278
16	4	4	13.1110	42.3420	**0.805951**	**4.58558**
17	1	5	20.5215	41.0219	0.243468	0.38314
18	2	5	23.6314	40.6483	0.165337	-0.08027
19	3	5	24.1979	42.8845	0.160233	0.17958
20	4	5	28.5071	43.7115	0.288632	0.09397

Table 11.1 Data and diagnostics for Example 1

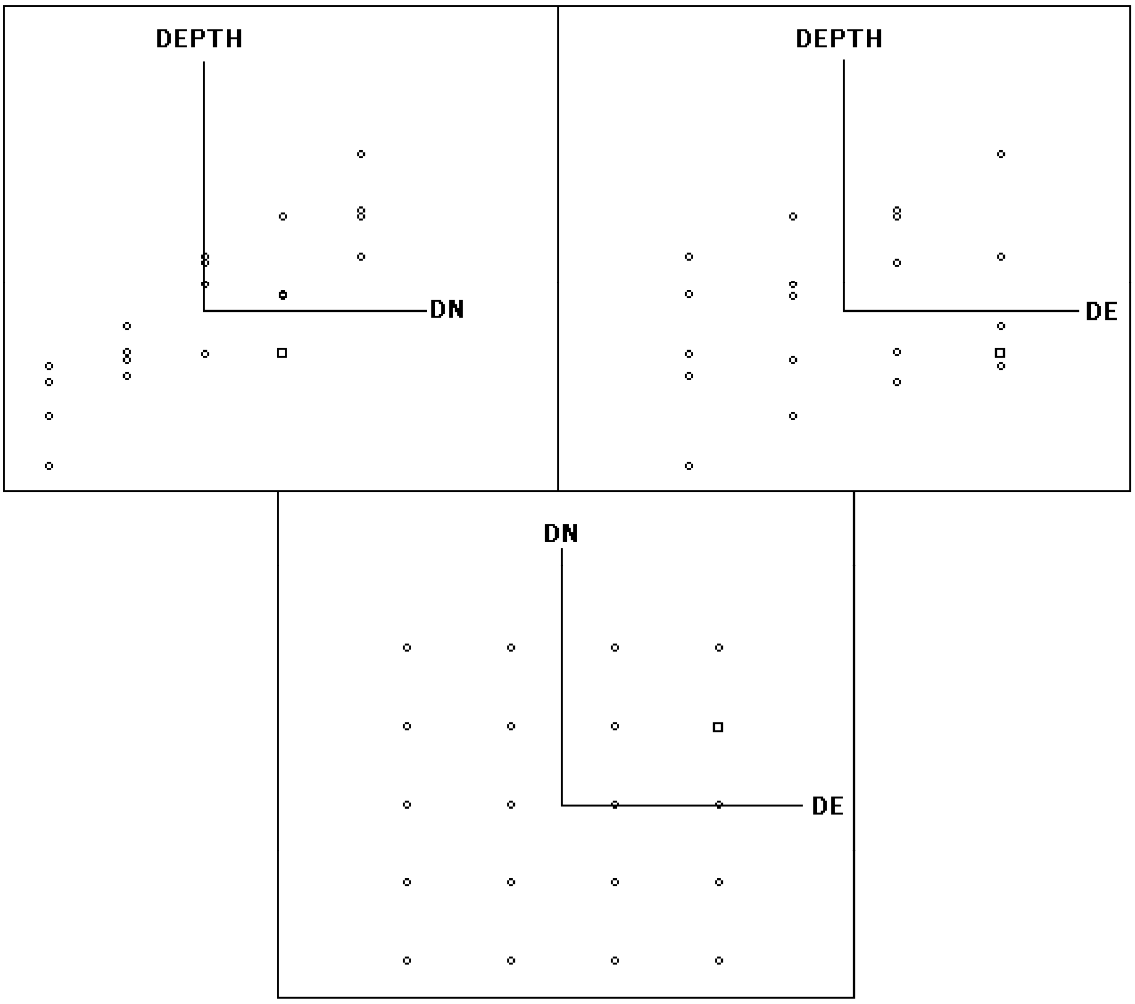


Figure 11.1 Scatterplot matrix for the 3 explanatory variables
(obs. 16 is shown as a square)

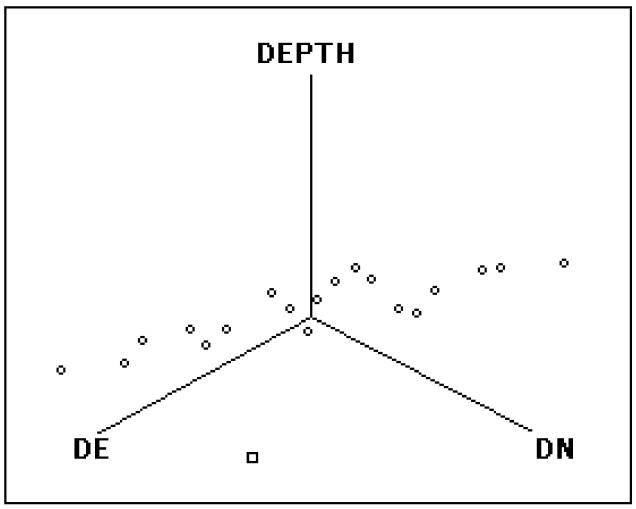


Figure 11.2 Rotated scatterplot showing the position of the high leverage point
(obs. 16 is shown as a square)

The depth value for observation 16 was set as a "typographical error", and should be 23.111 instead of 13.111. What does this error and resulting high leverage point do to a regression of concentration versus the three explanatory variables? From the t-ratios of table 11.2 it is seen that DN and perhaps DE appear to be significantly related to Conc, but that depth (D) is not. This is exactly opposite of what is known to be true.

Conc = 28.9 + 0.991 DE + 1.60 DN + 0.091 D

n = 20 s = 2.14 $R^2 = 0.71$

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	28.909	1.582	18.28	0.000
Slopes β_k				
DE	0.991	0.520	1.90	0.075
DN	1.596	0.751	2.13	0.049
D	0.091	0.186	0.49	0.632

Table 11.2 Regression statistics for Example 1

One outlier has had a severe detrimental effect on the regression coefficients and model structure. Points of high leverage and influence should always be examined before accepting a regression model, to determine if they represent errors. Suppose that the "typographical error" was detected and corrected. Table 11.3 shows that the resulting regression relationship is drastically changed:

C = 29.2 – 0.419 DE – 0.82 DN + 0.710 D

n = 20 s = 1.91 $R^2 = 0.77$

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	29.168	1.387	21.03	0.000
Slopes β_k				
DE	–0.419	0.833	–0.50	0.622
DN	–0.816	1.340	–0.61	0.551
D	0.710	0.339	2.10	0.052

Table 11.3 Regression statistics for the corrected Example 1 data

Based on the t-statistics, DE and DN are not significantly related to C, while depth is related. The intercept of 29 is close to the true value of 30, and the slope for depth (0.7) is not far from the true value of 0.5. For observation 16, hi = 0.19 and DFFITS = 0.48, both well below their critical values. Thus no observations have undue influence on the regression equation. Since

DE and DN do not appear to belong in the regression model, dropping them produces the equation of table 11.4, with values very close to the true values from which the data were generated. Thus by using regression diagnostics to inspect observations deemed unusual, a poor regression model was turned into an acceptable one.

Conc = 29.0 + 0.511 D

n = 20 s = 1.83 $R^2 = 0.77$

Parameter	Estimate	Std.Err(β)	t-ratio	p
Intercept β_0	29.036	1.198	24.23	0.000
Slope D	0.511	0.067	7.65	0.000

Table 11.4 Final regression model for the corrected Example 1 data

11.5.3 Multi-Collinearity

It is very important that practitioners of MLR understand the causes and consequences of multi-collinearity, and can diagnose its presence. Multi-collinearity is the condition where at least one explanatory variable is closely related to one or more other explanatory variables. It results in several undesirable consequences for the regression equation, including:

1) Equations acceptable in terms of overall F-tests may have slope coefficients with magnitudes which are unrealistically large, and whose partial F or t-tests are found to be insignificant.
2) Coefficients may be unrealistic in sign (a negative slope for a regression of streamflow vs. precipitation, (etc). Usually this occurs when two variables describing approximately the same thing are counter-balancing each other in the equation, having opposite signs.
3) Slope coefficients are unstable. A small change in one or a few data values could cause a large change in the coefficients.
4) Automatic procedures such as stepwise, forwards and backwards methods produce different models judged to be "best".

Concern over multi-collinearity should be strongest when the purpose is to make inferences about coefficients. Concern can be somewhat less when only predictions are of interest, provided that these predictions are for cases within the observed range of the x data.

An excellent diagnostic for measuring multi-collinearity is the variance inflation factor (VIF) presented by Marquardt (1970). For variable j the VIF is

$$VIF_j = 1/(1-R_j^2) \qquad [11.6]$$

where R_j^2 is the R^2 from a regression of the jth explanatory variable on all of the other explanatory variables -- the equation used for adjustment of x_j in partial plots. The ideal is VIF_j

$\cong 1$, corresponding to $R_j^2 \cong 0$. Serious problems are indicated when $VIF_j > 10$ ($R_j^2 > 0.9$). A useful interpretation of VIF is that multi-collinearity "inflates" the width of the confidence interval for the jth regression coefficient by the amount $\sqrt{VIF_j}$ compared to what it would be with a perfectly independent set of explanatory variables.

11.5.3.1 Solutions for multi-collinearity

There are four options for working with a regression equation having one or more high VIF values.

1) **Center the data**. A simple solution which works in some specific cases is to center the data. Multi-collinearity can arise when some of the explanatory variables are functions of other explanatory variables, such as for a polynomial regression of y against x and x^2. When x is always of one sign, there may be a strong relationship between it and its square. Centering redefines the explanatory variables by subtracting a constant from the original variable, and then recomputing the derived variables. This constant should be one which produces about as many positive values as negative values, such as the mean or median. When all of the derived explanatory variables are recomputed as functions (squares, products, etc.) of these centered variables, their multi-collinearity will be reduced.

   Centering is a mathematical solution to a mathematical problem. It will not reduce multi-collinearity between two variables which are not mathematically derived one from another. It is particularly useful when the original explanatory variable has been defined with respect to some arbitrary datum (time, distance, temperature) and is easily fixed by resetting the datum to roughly the middle of the data. In some cases the multi-collinearity can be so severe that the numerical methods used by the statistical software fail to perform the necessary matrix computations correctly. Such numerical problems occur frequently when doing trend surface analysis (e.g., fitting a high order polynomial of distances north of the equator and west of Greenwich) or trend analysis (e.g., values are a polynomial of years). This will be demonstrated in Example 2.

2) **Eliminate variables**. In some cases prior judgment suggests the use of several different variables which describe related but not identical attributes. Examples of this might be: air temperature and dew point temperature, the maximum 1-hour rainfall, and the maximum 2-hour rainfall, river basin population and area in urban land use, basin area forested and basin area above 6,000 feet elevation, and so on. Such variables may be strongly related as shown by their VIFs, and one of them must eliminated on judgmental grounds, or on the basis of comparisons of models fit with one eliminated versus the other eliminated, in order to lower the VIF.

3) **Collect additional data**. Multi-collinearity can sometimes be solved with only a few additional but strategically selected observations. Suppose some attributes of river basins are being studied, and small basins tend to be heavily forested while large basins tend to be less heavily forested. Discerning the relative importance of size versus the importance of forest cover will prove to be difficult. Strong multi-collinearity will result from including both variables in the regression equation. To solve this and allow the effects of each variable to be judged separately, collect additional samples from a few small less forested basins and a few large but heavily-forested basins. This produces a much more reliable model. Similar problems arise in ground-water quality studies, where rural wells are shallow and urban wells are deeper. Depth and population density may show strong multi-collinearity, requiring some shallow urban and deeper rural wells to be sampled.

4) **Perform ridge regression.** Ridge regression was proposed by Hoerl and Kenard (1970). Montgomery and Peck (1982) give a good brief discussion of it. It is based on the idea that the variance of the slope estimates can be greatly reduced by introducing some bias into them. It is a controversial but useful method in multiple regression.

<u>Example 2 -- centering</u>

The natural log of concentration of some contaminant in a shallow groundwater plume is to be related to distance east and distance north of a city. The city was arbitrarily chosen as a geographic datum. The data are presented in table 11.5.

Since the square of distance east (DESQ) must be strongly related to DE, and similarly DNSQ and DN, and DE•DN with both DE and DN, multi-collinearity between these variables will be detected by their VIFs. Using the rule that any VIF above 10 indicates a strong dependence between variables, table 11.6 shows that all variables have high VIFs. Therefore all of the slope coefficients are unstable, and no conclusions can be drawn from the value of 10.5 for DE, or 15.1 for DN, etc. This cannot be considered a good regression model, even though the R^2 is large.

Obs. #	C	ln(C)	DE	DN	DESQ	DNSQ	DE•DN
1	14	2.63906	17	48	289	2304	816
2	88	4.47734	19	48	361	2304	912
3	249	5.51745	21	48	441	2304	1008
4	14	2.63906	23	48	529	2304	1104
5	29	3.36730	17	49	289	2401	833
6	147	4.99043	19	49	361	2401	931
7	195	5.27300	21	49	441	2401	1029
8	28	3.33220	23	49	529	2401	1127
9	21	3.04452	17	50	289	2500	850
10	276	5.62040	19	50	361	2500	950
11	219	5.38907	21	50	441	2500	1050
12	40	3.68888	23	50	529	2500	1150
13	22	3.09104	17	51	289	2601	867
14	234	5.45532	19	51	361	2601	969
15	203	5.31320	21	51	441	2601	1071
16	35	3.55535	23	51	529	2601	1173
17	15	2.70805	17	52	289	2704	884
18	115	4.74493	19	52	361	2704	988
19	180	5.19296	21	52	441	2704	1092
20	16	2.77259	23	52	529	2704	1196

Table 11.5 Data for Example 2

DE and DN are centered by subtracting their medians. Following this, the three derived variables DESQ, DNSQ and DEDN are recomputed, and the regression rerun. Table 11.7 give the results, showing that all multi-collinearity is completely removed. The coefficients for DE and DN are now more reasonable in size, while the coefficients for the derived variables are exactly the same. The t-statistics for DE and DN have changed because their uncentered values were unstable and t-tests unreliable. Note that the s and R^2 are also unchanged. In fact, this is exactly the same model as the uncentered equation, but only in a different and centered coordinate system.

$\ln(C) = -479 + 10.5\,DE + 15.1\,DN - 0.264\,DESQ - 0.151\,DNSQ + 0.0014\,DEDN$

$n = 20$ $s = 0.27$ $R^2 = 0.96$

Parameter	Estimate	Std.Err(β)	t-ratio	p	VIF
Intercept β_0	−479.03	91.66	−5.23	0.000	
Slopes β_k					
DE	10.55	1.12	9.40	0.000	1751.0
DN	15.14	3.60	4.20	0.001	7223.9
DESQ	−0.26	0.015	−17.63	0.000	501.0
DNSQ	−0.15	0.04	−4.23	0.001	7143.9
DEDN	0.001	0.02	0.07	0.943	1331.0

Table 11.6 Regression statistics and VIFs for Example 2

$\ln(C) = 5.76 + 0.048\,DE + 0.019\,DN - 0.264\,DESQ - 0.151\,DNSQ + 0.001\,DNDE$

$n = 20$ $s = 0.27$ $R^2 = 0.96$

Parameter	Estimate	Std.Err(β)	t-ratio	p	VIF
Intercept β_0	5.76	0.120	48.15	0.000	
Slopes β_k					
DE	0.048	0.027	1.80	0.094	1.0
DN	0.019	0.042	0.44	0.668	1.0
DESQ	−0.264	0.015	−17.63	0.000	1.0
DNSQ	−0.151	0.036	−4.23	0.001	1.0
DEDN	0.001	0.019	0.07	0.943	1.0

Table 11.7 Regression statistics and VIFs for centered Example 2 data

11.6 Choosing the Best MLR Model

One of the major issues in multiple regression is the appropriate approach to variable selection. The benefit of adding additional variables to a multiple regression model is to account for or explain more of the variance of the response variable. The cost of adding additional variables is that the degrees of freedom decreases, making it more difficult to find significance in hypothesis tests and increasing the width of confidence intervals. A good model will explain as much of the variance of y as possible with a small number of explanatory variables.

The first step is to consider only explanatory variables which ought to have some effect on the dependent variable. There must be plausible theory behind why a variable might be expected to influence the magnitude of y. Simply minimizing the SSE or maximizing R^2 are not sufficient criteria. In fact, any explanatory variable will reduce the SSE and increase the R^2 by some small amount, even those irrelevant to the situation (or even random numbers!). The benefit of

adding these unrelated variables, however, is small compared to the cost of a degree of freedom. Therefore the choice of whether to add a variable is based on a "cost-benefit analysis", and variables enter the model only if they make a significant improvement in the model. There are at least two types of approaches for evaluating whether a new variable sufficiently improves the model. The first approach uses partial F or t-tests, and when automated is often called a "stepwise" procedure. The second approach uses some overall measure of model quality. The latter has many advantages.

11.6.1 Stepwise Procedures

Stepwise procedures are automated model selection methods in which the computer algorithm determines which model is preferred. There are three versions, usually called forwards, backwards, and stepwise. These procedures use a sequence of partial F or t-tests to evaluate the significance of a variable. The three versions do not always agree on a "best" model, especially when multi-collinearity is present. They also do not evaluate all possible models, and so cannot guarantee that the "best" model is even tested. They were developed prior to modern computer technology, taking shortcuts to avoid running all possible regression equations for comparison. Such shortcuts are no longer necessary.

Forward selection starts with only an intercept and adds variables to the equation one at a time. Once in, each variable stays in the model. All variables not in the model are evaluated with partial F or t statistics in comparison to the existing model. The variable with the highest significant partial F or t statistic is included, and the process repeats until either all available variables are included or no new variables are significant. One drawback to this method is that the resulting model may have coefficients which are not significantly different from zero; they must only be significant when they enter. A second drawback is that two variables which each individually provide little explanation of y may never enter, but together the variables would explain a great deal. Forward selection is unable to capitalize on this situation.

Backward elimination starts with all explanatory variables in the model and eliminates the one with the lowest partial-F statistic (lowest $|t|$). It stops when all remaining variables are significant. The backwards algorithm does ensure that the final model has only significant variables, but does not ensure a "best" model because it also cannot consider the combined significance of groups of variables.

Stepwise regression combines the ideas of forward and backward. It alternates between adding and removing variables, checking significance of individual variables within and outside the model. Variables significant when entering the model will be eliminated if later they test as insignificant. Even so, stepwise does not test all possible regression models.

Example 3:

Haan (1977) attempted to relate the mean annual runoff of several streams (ROFF) with 9 other variables: the precipitation falling at the gage (PCIP), the drainage area of the basin (AREA), the average slope of the basin (SLOPE), the length of the drainage basin (LEN), the perimeter of the basin (PERIM), the diameter of the largest circle which could be inscribed within the drainage basin (DI), the "shape factor" of the basin (Rs), the stream frequency -- the ratio of the number of streams in the basin to the basin area (FREQ), and the relief ratio for the basin (Rr). The data are found in Appendix C14. Haan chose to select a 3-variable model (using PCIP, PERIM and Rr) based on a levelling off of the incremental increase in R^2 as more variables were added to the equation (see figure 11.3).

What models would be selected if the stepwise or overall methods are applied to this data? If a forwards routine is performed, no single variables are found significant at $\alpha = 0.05$, so an intercept-only model is declared "best". Relaxing the entry criteria to a larger α, AREA is first entered into the equation. Then Rr, PCIP, and PERIM are entered in that order. Note that AREA has relatively low significance once the other three variables are added to the model (Model 4).

Forwards		Model 1	Model 2	Model 3	Model 4
AREA	β	0.43	0.81	0.83	−0.62
	t	1.77	4.36	4.97	−1.68
Rr	β		0.013	0.011	0.009
	t		3.95	3.49	4.89
PCIP	β			0.26	0.54
	t			1.91	5.05
PERIM	β				1.02
	t				4.09

The backwards model begins with all variables in the model. It checks all partial t or F statistics, throwing away the variable which is least significant. Here the least significant single variable is AREA. So while forwards made AREA the first variable to bring in, backwards discarded AREA first of all! Then other variables were also removed, resulting in a model with Rr, PCIP, PERIM, DI and FREQ remaining in the model. Multi-collinearity between measures of drainage basin size, as well as between other variables, has produced models from backwards and forwards procedures which are quite different from each other. The slope coefficient for DI is also negative, suggesting that runoff decreases as basin size increases. Obviously DI is counteracting another size variable in the model (PERIM) whose coefficient is large.

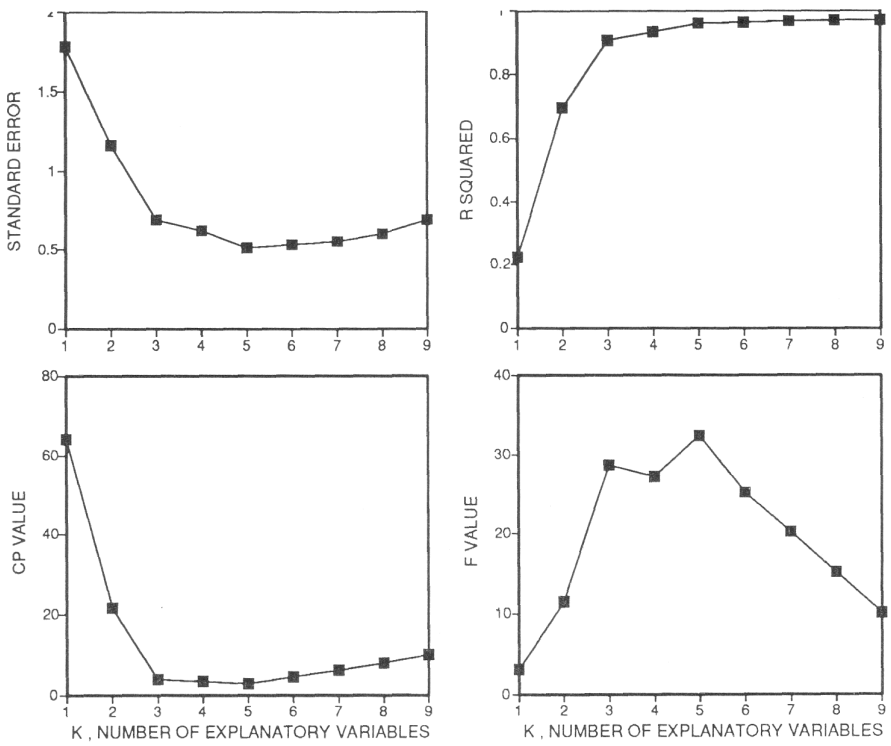


Figure 11.3 Magnitude of σ^2, R^2, Cp and F as a function of the number of explanatory variables, for the best k explanatory variable model.

Stepwise first enters AREA, Rr, PCIP and PERIM. At that point, the t-value for AREA drops from near 5 to −1.6, so AREA is dropped from the model. DI and FREQ are then entered, so that stepwise results in the same 5-variable model as did backwards.

Stepwise		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
AREA	β	0.43	0.81	0.83	−0.62			
	t	1.77	4.36	4.97	−1.68			
Rr	β		0.013	0.011	0.009	0.010	0.010	0.011
	t		3.95	3.49	4.89	5.19	5.02	6.40
PCIP	β			0.260	0.539	0.430	0.495	0.516
	t			1.91	5.05	4.62	5.39	6.71
PERIM	β				1.021	0.617	0.770	0.878
	t				4.09	8.24	6.98	8.38
DI	β						−1.18	−1.30
	t						−1.75	−2.32
FREQ	β							0.36
	t							2.14

11.6.2 Overall Measures of Quality

Three newer statistics can be used to evaluate each of the 2^k regressions equations possible from k candidate explanatory variables. These are Mallow's Cp, the PRESS statistic, and the adjusted R^2.

Mallow's Cp, is designed to achieve a good compromise between the desire to explain as much variance in y as possible (minimize bias) by including all relevant variables, and to minimize the variance of the resulting estimates (minimize the standard error) by keeping the number of coefficients small. The Cp statistic is

$$Cp = p + \frac{(n-p) \bullet (s_p^2 - \hat{\sigma}^2)}{\hat{\sigma}^2} \qquad [11.7]$$

where n is the number of observations, p is the number of coefficients (number of explanatory variables plus 1), s_p^2 is the mean square error (MSE) of this p coefficient model, and $\hat{\sigma}^2$ is the best estimate of the true error, which is usually taken to be the minimum MSE among the 2^k possible models. The best model is the one with the lowest Cp value. When several models have nearly equal Cp values, they may be compared in terms of reasonableness, multi-collinearity, importance of high influence points, and cost in order to select the model with the best overall properties.

The second overall measure is the PRESS statistic. PRESS was defined in Chapter 9 as the sum of the squared prediction errors $e_{(i)}$. By minimizing PRESS, the model with the least error in the prediction of future observations is selected. PRESS and Cp generally agree as to which model is "best", even though their criteria for selection are not identical.

A third overall measure is the adjusted R^2 (R^2_a). This is an R^2 value adjusted for the number of explanatory variables (or equivalently, the degrees of freedom) in the model. The model with the highest R^2_a is identical to the one with the smallest standard error (s) or its square, the mean squared error (MSE). To see this, in Chapter 9 R^2 was defined as a function of the total (SS_y) and error (SSE) sum of squares for the regression model:

$$R^2 = 1 - (SSE / SS_y) \qquad [11.8]$$

The weakness of R^2 is that it must increase, and the SSE decrease, when any additional variable is added to the regression. This happens no matter how little explanatory power that variable has. R^2_a is adjusted to offset the loss in degrees of freedom by including as a weight the ratio of total to error degrees of freedom:

$$R^2_a = 1 - \frac{(n-1)}{(n-p)} \frac{SSE}{SS_y} = 1 - \frac{MSE}{(SS_y/(n-1))} \qquad [11.9]$$

As $(SS_y/(n-1))$ is constant for a given data set, R^2_a increases as MSE decreases. Either maximize R^2_a or minimize MSE as an overall measure of quality. However, when p is considerably smaller than n, R^2_a is a less sensitive measure than either PRESS or Cp. PRESS has the additional advantage of being a validation criteria.

Overall methods use the computer to perform large computations (such as Cp and PRESS for many models), letting the scientist judge which model to use. This allows flexibility in choosing between models. For example, two "best" models may be nearly identical in terms of their Cp, R^2_a and/or PRESS statistics, yet one involves variables that are much less expensive to measure than the other. The less expensive model can be selected with confidence. In contrast, stepwise procedures ask the computer to judge which model is best. Their combination of inflexible criteria and inability to test all models often results in the selection of something much less than the best model.

<u>Example 3, continued</u>
Instead of the stepwise procedures run on Haan's data, models are evaluated using the overall statistics Cp and PRESS. Smaller values of Cp and PRESS are associated with better models. Computing PRESS and Cp for the $2^9 = 512$ possible regression models can be done with modern statistical software. A list of these statistics for the two best k-variable models, where best is defined as the highest R^2, is given in table 11.8.

Based on Cp, the best model would be the 5 variable model having PCIP, PERIM, DI, FREQ and Rr as explanatory variables -- the same model as selected by stepwise and forwards. Remember that there is no guarantee that stepwise procedures regularly select the lowest Cp or PRESS models. The advantage of using an overall statistic like Cp is that options are given to the scientist to select what is best. If the modest multi-collinearity (VIF=5.1) between PERIM and DI is of concern, with its resultant negative slope for DI, the model with the next lowest Cp that does not contain both these variables (a four-variable model with Cp= 3.6) could be selected. If the scientist decided AREA must be in the model, the lowest CP model containing AREA (the same four-variable model) could be selected. Cp and PRESS allow model choice to be based on multiple criteria such as prediction quality (PRESS), low VIF, cost, etc..

# of Vars	R-sq	PRESS	C-p	Max VIF	PCIP	AREA	SLOPE	LEN	PERIM	DIs	RR	FREQ	Rr
					X: Variable is in the model								
1	22.2	47.3	64.2	---		X							
1	21.6	49.3	64.7	---					X				
2	69.6	20.6	21.6	1.4		X							X
2	68.1	24.7	23.0	1.4					X				X
3	90.5	10.0	3.9	1.5	X				X				X
3	80.7	19.3	13.2	1.2	X		X		X				
4	93.2	8.0	3.4	3.9	X				X	X			X
4	93.0	7.7	3.6	19.5	X	X			X				X
5	95.9	6.9	2.9	5.1	X				X	X		X	X
5	94.7	7.2	4.0	21.6	X	X	X		X				X
6	96.2	6.3	4.6	19.6	X				X	X	X	X	X
6	96.1	10.0	4.7	8.2	X		X		X	X		X	X
7	96.6	7.5	6.2	21.1	X		X		X	X	X	X	X
7	96.6	6.4	6.2	127.9	X	X			X	X	X	X	X
8	96.8	53.7	8.0	23.4	X		X	X	X	X	X	X	X
8	96.7	10.2	8.1	200.8	X	X	X		X	X	X	X	X
9	96.8	59.7	10.0	227.5	X	X	X	X	X	X	X	X	X

Table 11.8 Statistics for several multiple regression models of Haan's data

11.7 Summary of Model Selection Criteria

Rules for selection of linear regression models are summarized in the 5 steps below:

1) **Should y be transformed?** To decide whether to transform the y variable, plot residuals versus predicted values for the untransformed data. Compare this to a residuals plot for the best transformed model, looking for three things:

1) constant variance across the range of $\hat{y}$,
2) normality of residuals, and
3) a linear pattern, not curvature.

The statistics R^2, SSE, Cp, and PRESS are <u>not</u> appropriate for comparison of models having different units of y.

2) **Should x (or several x's) be transformed?** Transformation of an x variable should be made using partial plots. Check for the same three patterns of constant variance, normality and

linearity. Considerable help can be obtained from statistics such as R^2 (maximize it), or SSE or PRESS (minimize it). Many transformations can be rapidly checked with such statistics, but a residuals plot should always be inspected prior to making any final decision.

3) **Which of several models, each with the same y and with the same number of explanatory variables, is preferable?** Use of R^2, SSE, or PRESS is appropriate here, but back it up with a residuals plot.

4) **Which of several nested models, each with the same y, is preferable?** Use the partial F test between any pair of nested models to find which is best. One may also select the model based on minimum Cp or minimum PRESS.

5) **Which of several models is preferable when each uses the same y variable but are not necessarily nested?** Cp or PRESS must be used in this situation.

11.8 Analysis of Covariance

Often there are factors which influence the dependent variable which are not appropriately expressed as a continuous variable. Examples of such grouped or qualitative variables include location (stations, aquifers, positions in a cross section), or time (day & night; winter & summer; before & after some event such as a flood, a drought, operation of a sewage treatment plant or reservoir). These factors are perfectly valid explanatory variables in a multiple regression context. They can be incorporated by the use of binary or "dummy" variables, essentially blending regression and analysis of variance into an analysis of covariance.

11.8.1 Use of One Binary Variable

To the simple one-variable regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad [11.10]$$

(again with subscripts i assumed), an additional factor is believed to have an important influence on Y for any given value of X. Perhaps this factor is a seasonal one: cold season versus warm season -- where some precise definition exists to classify all observations as either cold or warm. A second variable, a binary variable Z, is added to the equation where

$$Z_i = \begin{cases} 0 \text{ if i is from cold season} \\ 1 \text{ if i is from warm season} \end{cases}$$

to produce the model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon. \qquad [11.11]$$

When the slope coefficient β_2 is significant, model 11.11 would be prefered to the SLR model 11.10. This also says that the relationship between Y and X is affected by season.

Consider H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$. The null hypothesis is tested using a student's t-test with (n−3) degrees of freedom. There are (n−3) because there are 3 betas being estimated. If the partial $|t| \geq t_{\alpha/2}$, H_0 is rejected, inferring that there are two models:

$\hat{Y} = b_0 + b_1 X$ for the cold season ($Z = 0$), and

$\hat{Y} = b_0 + b_1 X + b_2$ for the warm season ($Z = 1$), or

$= (b_0 + b_2) + b_1 X$.

Therefore the regression lines differ for the two seasons. Both seasons have the same slope, but different intercepts, and will plot as two parallel lines (figure 11.4).

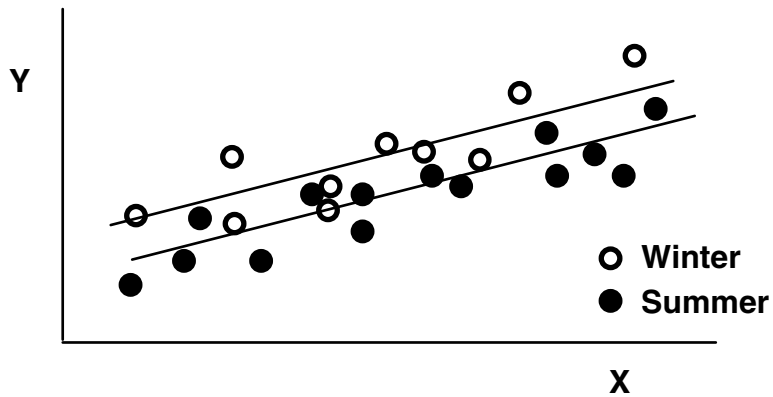


Figure 11.4 Regression lines for data differing in intercept between two seasons

Suppose that the relationship between X and Y for the two seasons is suspected not only to differ in intercept, but in slope as well. Such a model is written as:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 Z X + \varepsilon \qquad [11.12]$$

or $$Y = (\beta_0 + \beta_2 Z) + (\beta_1 + \beta_3 Z) \bullet X + \varepsilon$$

The intercept equals β_0 for the cold season and $\beta_0 + \beta_2$ for the warm season; the slope equals β_1 for the cold season and $\beta_1 + \beta_3$ for the warm season. This model is referred to as an "interaction model" because of the use of the explanatory variable $Z X$, the interaction (product) of the original predictor X and the binary variable Z.

To determine whether the simple regression model with no Z terms can be improved upon by model 11.12, the following hypotheses are tested:

H_0 : $\beta_2 = \beta_3 = 0$ versus H_1 : β_2 and/or $\beta_3 \neq 0$.

A nested F statistic is computed $$F = \frac{(SSE_s - SSE_c) / (df_s - df_c)}{(SSE_c / df_c)}$$

where s refers to the simpler (no Z terms) model, and c refers to the more complex model. For the two nested models 11.10 and 11.12 this becomes

$$F = \frac{(SSE_{11.10} - SSE_{11.12}) / 2}{MSE_{11.12}}$$

where $MSE_{11.12} = SSE_{11.12} /(n-4)$, rejecting H_0 if $F > F_{\alpha,2,\, n-4}$.

If H_0 is rejected, model 11.12 should also be compared to model 11.11 (the shift in intercept only model) to determine whether there is a change in slope in addition to the change in intercept, or whether the rejection of model 11.10 in favor of 11.12 was due only to a shift in intercept. The null hypothesis H_0': $\beta_3 = 0$ is compared to H_1': $\beta_3 \neq 0$ using the test statistic

$$F = \frac{(SSE_{11.11} - SSE_{11.12}) / 1}{MSE_{11.12}}$$

rejecting H_0' if $F > F_{\alpha,1,\, n-4}$.

Assuming H_0 and H_0' are both rejected, the model can be expressed as the two separate equations (see figure 11.5):

$\hat{Y} = b_0 + b_1 X$ cold season

$\hat{Y} = (b_0 + b_2) + (b_1 + b_3) X$ warm season

Furthermore, the coefficient values in these two equations will be exactly those computed if the two regressions were estimated by separating the data, and computing two separate regression equations. By using analysis of covariance, however, the significance of the difference between those two equations has been established.

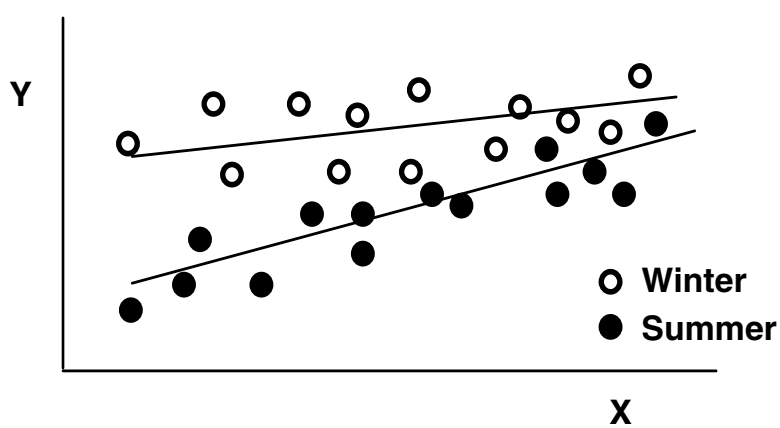


Figure 11.5 Regression lines differing in slope and intercept for data from two seasons

11.8.2 Multiple Binary Variables

In some cases, the factor of interest must be expressed as more than two categories: 4 seasons, 12 months, 5 stations, 3 flow conditions (rising limb, falling limb, base flow), etc. To illustrate, assume there are precise definitions of 3 flow conditions so that all discharge (X_i) and

concentration (Y_i) pairs are classified as either rising, falling, or base flow. Two binary variables are required to express these three categories -- there is always one less binary variable required than the number of categories.

$$\text{Let } R_i = \begin{cases} 1 \text{ if i is a rising limb observation} \\ 0 \text{ otherwise} \end{cases}$$

$$\text{Let } D_i = \begin{cases} 1 \text{ if i is falling limb observation} \\ 0 \text{ otherwise} \end{cases}$$

so that

category	value of R	value of D
rising	1	0
falling	0	1
base flow	0	0

The following model results:

$$Y = \beta_0 + \beta_1 X + \beta_2 R + \beta_3 D + \varepsilon \qquad [11.13]$$

To test H_0: $\beta_2 = \beta_3 = 0$ versus H_1: β_2 and/or $\beta_3 \neq 0$, F tests comparing simpler and more complex models are again performed. To compare model 11.13 versus the SLR model 11.10 with no rising or falling terms,

$$F = \frac{(SSE_{11.10} - SSE_{11.13}) / 2}{MSE_{11.13}} \qquad \text{where } MSE_{11.13} = SSE_{11.13} / (n-4),$$

rejecting H_0 if $F > F_{2,\, n-4,\, \alpha}$.

To test for differences between each pair of categories:

1. Is rising different from base flow? This is tested using the t-statistic on the coefficient β_2. If $|t| > t_{\alpha/2}$ on n–4 degrees of freedom, reject H_0 where H_0: $\beta_2 = 0$.

2. Is falling different from base flow? This is tested using the t-statistic on the coefficient β_3. If $|t| > t_{\alpha/2}$ with n–4 degrees of freedom, reject H_0 where H_0: $\beta_3 = 0$.

3. Is rising different from falling? There are two ways to determine this.
   (a) the standard error of the difference ($b_2 - b_3$) must be known. The null hypothesis is H_0: $(\beta_2 - \beta_3) = 0$. The estimated variance of $b_2 - b_3$,
   $$Var(b_2 - b_3) = Var(b_2) + Var(b_3) - 2Cov(b_2, b_3)$$
   where Cov is the covariance between b_2 and b_3. To determine these terms, the matrix $(X'X)^{-1}$ and s^2 (s^2 is the mean square error) are required. Then

$\hat{Var}(b_2) = C_{22} \bullet s^2$, $\hat{Var}(b_3) = C_{33} \bullet s^2$, and $\hat{Cov}(b_2,b_3) = C_{23} \bullet s^2$.
The test statistic is $t = (b_2 - b_3)/\sqrt{Var\ (b_2 - b_3)}$. If $|t| > t_{\alpha/2}$ with n–4 degrees of freedom, reject H_0.

(b) The binary variables can be re-defined so that a direct contrast between rising and falling is possible. This occurs when either is set as the (0,0) "default" case. This will give answers identical to (a).

Ever greater complexity can be added to these kinds of models, using multiple binary variables and interaction terms such as

$$Y = \beta_0 + \beta_1 X + \beta_2 R + \beta_3 D + \beta_4 R X + \beta_5 D X + \varepsilon. \qquad [11.14]$$

The procedures for selecting models follow the pattern described above. The significance of an individual β coefficient, given all the other βs, can be determined from the t statistic. The comparison of two models, where the set of explanatory variables for one model is a subset of those used in the other model, is computed by a nested F test. The determination of whether two coefficients in a given model differ significantly from each other is computed either by re-defining the variables, or by using a t test after estimating the variance of the difference between the coefficients based on the elements of the $(X'X)^{-1}$ matrix and s^2.

Exercises

11.1 In Appendix C15 are data from 42 small urban drainage basins located in several cities around the United States (Mustard et al., 1987). The dependent variable is the log of the total nitrogen load for the basin -- the y transformation decision has already been made. There are eight possible explanatory variables to use for prediction purposes. The definitions of all nine variables are as follows.

LOGTN	log total nitrogen load
LOGCA	log contributing area
LOGIMP	log impervious area
MMJTEMP	mean minimum January temperature
MSRAIN	mean seasonal rainfall
PRES	percentage of area residential
PNON	percentage of area non-urban
PCOMM	percentage of area commercial
PIND	percentage of area industrial

Don't bother with transformations of the x variables either -- use these variables as they are. Pay special attention to multi-collinearity. Try one or more of the approaches described in this chapter to select the best model for predicting LOGTN from these explanatory variables.

11.2 Analysis of Covariance. The following 10 possible models describe the variation in sand-size particles (0.125 – 0.250 mm) in the Colorado River at Lees Ferry, AZ. Select the best model from this set of 10 and interpret its meaning. The basic model describes a quadratic relationship between concentration and discharge (X). Do intercept and/or slope vary with the three seasons (summer S, winter W, or other)? Use $\alpha = 0.05$ for all hypothesis tests.

Basic model $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

where $Y = \ln$ (concentration of suspended sands)

$X = \ln$ (discharge)

	Month											
Binary Variables	1	2	3	4	5	6	7	8	9	10	11	12
S	0	0	0	0	0	0	1	1	1	1	0	0
W	1	1	0	0	0	0	0	0	0	0	1	1

Model #	Explanatory variables	SSE	df(error)
1	X, X^2	69.89	124
2	X, X^2, S	65.80	123
3	X, X^2, S, SX	65.18	122
4	X, X^2, S, SX, SX^2	64.84	121
5	X, X^2, W	63.75	123
6	X, X^2, W, WX	63.53	122
7	X, X^2, W, WX, WX^2	63.46	121
8	X, X^2, S, W	63.03	122
9	X, X^2, S, W, SX, WX	62.54	120
10	X, X^2, S, W, SX, WX, SX^2, WX^2	61.45	118

11.3 The Ogallala aquifer was investigated to determine relationships between uranium and other concentrations in its waters. Construct a regression model to relate uranium to total dissolved solids and bicarbonate, using the data in Appendix C16. What is the significance of these predictor variables?

11.4 You are asked to estimate uranium concentrations in irrigation waters from the Ogallala aquifer for a local area. Four supply wells pump waters with the characteristics given below. The relative amounts of water pumped by each well are also given below. Using this and the regression equation of Exercise 11.3, estimate the mean concentration of uranium in the water applied to this area.

Well #	Relative volume of water used	TDS	Bicarbonate
1	2	500	≤ 50%
2	1	900	≤ 50%
3	1	400	> 50%
4	2	600	> 50%

Chapter 12

Trend Analysis

Concentrations and loads of phosphorus have been observed at numerous tributaries to an important estuary over a 20-year period. Have concentrations and/or loads changed over time? Have concentrations changed when changing flow conditions are taken into account (the early years were during a very dry period), or are all changes simply due to more precipitation in the latter years? Is there an observable effect associated with a ban on phosphorus compounds in detergents which was implemented in the middle of the period of record?

Groundwater levels were recorded for many wells in a study area over 14 years. During the ninth year development of the area increased withdrawals dramatically. Is there evidence of decreasing water levels in the region's wells after versus before the increased pumpage?

Benthic invertebrate and fish population data were collected at twenty stations along one hundred miles of a major river. Do these data change in a consistent manner going downstream? What is the overall rate of change in population numbers over the one hundred miles?

Procedures for trend analysis build on those in previous chapters on regression and hypothesis testing. The explanatory variable of interest is usually time, though spatial or directional trends (such as downstream order or distance downdip) may also be investigated. Tests for trend have been of keen interest in environmental sciences over the last 10-15 years. Detection of both sudden and gradual trends over time with and without adjustment for the effects of confounding variables have been employed. In this chapter the various tests are classified, and their strengths and weaknesses compared.

12.1 General Structure of Trend Tests

12.1.1 Purpose of Trend Testing

A series of observations of a random variable (concentration, unit well yield, biologic diversity, etc.) have been collected over some period of time. We would like to determine if their values generally increase or decrease (getting "better" or "worse"). In statistical terms this is a determination of whether the probability distribution from which they arise has changed over time. We would also like to describe the amount or rate of that change, in terms of changes in some central value of the distribution such as a mean or median. Interest may be in data at one location, or all across the country. Figure 12.1 presents an example of the results of trend tests for bacteria at sites throughout the United States.

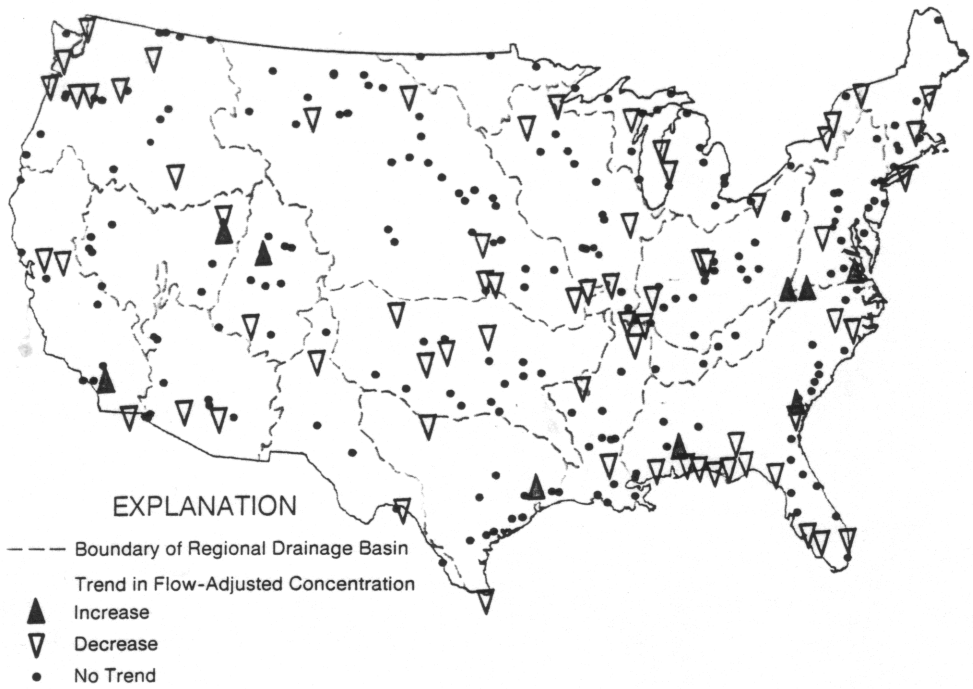


Figure 12.1 Trends in flow-adjusted concentrations of fecal streptococcus bacteria, 1974-1981 (from Smith et al., 1987).

The null hypothesis: H_0 is that there is no trend. However, any given test brings with it a precise mathematical definition of what is meant by "no trend", including a set of background assumptions usually related to type of distribution and serial correlation. The outcome of the test is a "decision" -- either H_0 is rejected or not rejected. Failing to reject H_0 does <u>not</u> mean that it was "proven" that there is no trend. Rather, it is a statement that the evidence available is not sufficient to conclude that there is a trend. Table 12.1 summarizes the possible outcomes of a statistical test in the context of trend analysis.

	True Situation	
Decision	No trend. H_0 true.	Trend exists. H_0 false.
Fail to reject H_0. "No trend"	Probability = $1-\alpha$	(Type II error) β
Reject H_0. "Trend"	(Type I error) significance level α	(Power) $1-\beta$

Table 12.1 Probabilities associated with possible outcomes of a trend test.
α = Prob (reject $H_0 | H_0$ true) and $1-\beta$ = Prob (reject $H_0 | H_0$ false)

The power ($1-\beta$) for the test can only be evaluated if the nature of the violation of H_0 that actually exists is known. This is never known in reality (if it were we wouldn't need a test), so a test must be found which has high power for the kind of data expected to be encountered. If a test is slightly more powerful in one instance but much less powerful than its alternatives in some other reasonable cases then it should not be used. The test selected should therefore be robust -- it should have relatively high power over all situations and types of data that might reasonably be expected to occur. Some of the characteristics commonly found in water resources data, and discussed in this chapter, are:

Distribution (normal, skewed, symmetric, heavy tailed)
Outliers (wild values that can't be shown to be measurement error)
Cycles (seasonal, weekly, tidal, diurnal)
Missing values (a few isolated values or large gaps)
Censored data (less-than values, historical floods)
Serial Correlation

12.1.2 Approaches to Trend Testing

Five types of trend tests are presented in table 12.2. They are classified based on two factors. The first, shown in the rows of the table, is whether the test is entirely parametric, entirely nonparametric, or a mixture of procedures. The second factor (columns) is whether there is some attempt to remove variation caused by other associated variables. The table uses the following notation:

Y = the random response variable of interest in the trend test,
X = an exogenous variable expected to affect the value of Y,
R = the residuals from a regression or LOWESS of Y versus X, and
T = time (often expressed in years).

Simple trend tests (not adjusted for X) are discussed in section 12.2. Tests adjusted for X are discussed in section 12.3.

	Not Adjusted for X	**Adjusted for X**
Nonparametric	Mann-Kendall trend test on Y	Mann-Kendall trend test on residuals R from LOWESS of Y on X
Mixed	----	Mann-Kendall trend test on residuals R from regression of Y on X
Parametric	Regression of Y on T	Regression of Y on X and T

Table 12.2 Classification of five types of tests for trend

If the trend is spatial rather than temporal, T will be downstream order, distance downdip, etc. Examples of X and Y include the following:

- For trends in surface water quality, Y would be concentration, X would be streamflow, and R would be called the flow-adjusted concentration;
- For trends in flood flows, Y would be streamflow, X would be the precipitation amount, and R would be called the precipitation-adjusted flow (the duration of precipitation used must be appropriate to the flow variable under consideration. For example, if Y is the annual flood peak from a 10 square mile basin then X might be the 1-hour maximum rainfall, whereas if Y is the annual flood peak for a 10,000 square mile basin then X might be the 24-hour maximum rainfall).
- For trends in groundwater levels, Y would be the change in monthly water level, X the monthly precipitation, and R would be called the precipitation-adjusted change in water level.

12.2 Trend Tests With No Exogenous Variable

12.2.1 Nonparametric Mann-Kendall Test

Mann (1945) first suggested using the test for significance of Kendall's tau where the X variable is time as a test for trend. This was directly analogous to regression, where the test for significance of the correlation coefficient r is also the significance test for a simple linear regression. The Mann-Kendall test can be stated most generally as a test for whether Y values tend to increase or decrease with T (monotonic change).

H_0: Prob $[Y_j > Y_i] = 0.5$, where time $T_j > T_i$.

H_1: Prob $[Y_j > Y_i] \neq 0.5$ (2-sided test).

No assumption of normality is required, but there must be no serial correlation for the resulting p-values to be correct. Typically the test is used for a more specific purpose -- to determine whether the central value or median changes over time. The spread of the distribution must remain constant, **though not necessarily in the original units**. If a monotonic transformation such as the ladder of powers is all that is required to produce constant variance, the test statistic will be identical to that for the original units. For example, in figure 12.2 a lognormal Y variable is plotted versus time. The variance of data around the trend is increasing. A Mann-Kendall test on Y has a p-value **identical** to that for the data of figure 12.3 -- the logarithms of the figure 12.2 data. The logs show an increasing median with constant variance. Only the central location changes. The Mann-Kendall test possesses the useful property of other nonparametric tests in that it is invariant to (monotonic) power transformations such as those of the ladder of powers. Since only the data **or any power transformation of the data** need be distributed similarly over T except for their central location in order to use the Mann-Kendall test, it is applicable in many situations.

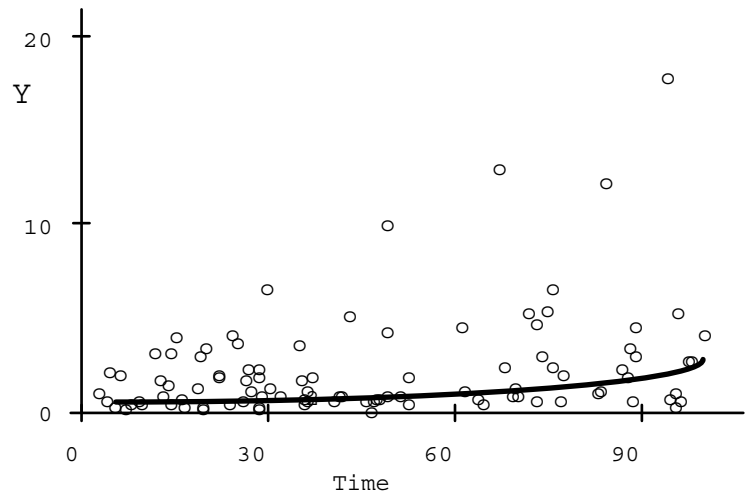


Figure 12.2 Y versus Time. Variance of Y increases over time.

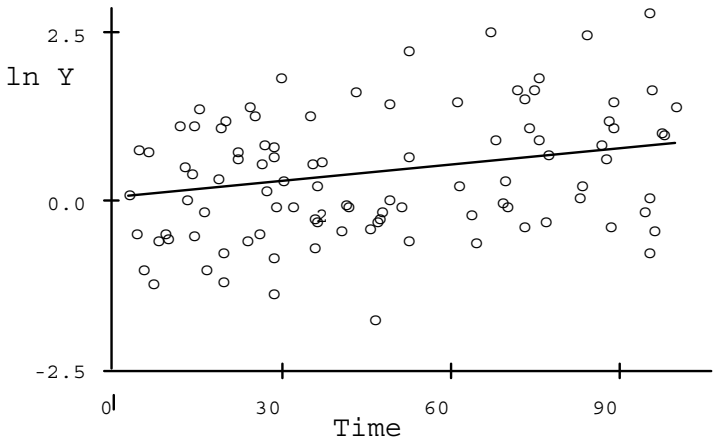


Figure 12.3 Logarithms of Y versus Time. The variance of Y is constant over time.

To perform the test, Kendall's S statistic is computed from the Y,T data pairs (see Chapter 8). The null hypothesis of no change is rejected when S (and therefore Kendall's τ of Y versus T) is significantly different from zero. We then conclude that there is a monotonic trend in Y over time.

An estimate of the rate of change in Y is also usually desired. If Y or some transformation of Y has a linear pattern versus T, the null hypothesis can be stated as a test for the slope coefficient $\beta_1 = 0$. β_1 is the rate of change in Y, or transformation of Y, over time.

12.2.2 Parametric Regression of Y on T

Simple linear regression of Y on T is a test for trend

$$Y = \beta_0 + \beta_1 \bullet T + \varepsilon$$

The null hypothesis is that the slope coefficient $\beta_1 = 0$. Regression makes stronger assumptions about the distribution of Y over time than does Mann-Kendall. It must be checked for normality of residuals, constant variance and linearity of the relationship (best done with residuals plots -- see Chapter 9). If Y is not linear over time, a transformation will likely be necessary. If all is ok, the t-statistic on b_1 is tested to determine if it is significantly different from 0. If the slope is nonzero, the null hypothesis of zero slope over time is rejected, and we conclude that there is a linear trend in Y over time. Unlike Mann-Kendall, the test results for regression will not be the same before and after a transformation of Y.

12.2.3 Comparison of Simple Tests for Trend

If the model form specified in a regression equation were known to be correct (Y is linear with T) and the residuals were truly normal, then fully-parametric regression would be optimal (most powerful and lowest error variance for the slope). Of course we can never know this in any real world situation. If the actual situation departs, even to a small extent, from these assumptions then the Mann-Kendall procedures will perform either as well or better (see Chapter 10, and Hirsch et. al., 1991, p.805-806).

There are practical cases where the regression approach is preferable, particularly in the multiple regression context (see next section). A good deal of care needs to be taken to insure it is correctly applied and to demonstrate that to the audience. When one is forced, by the sheer number of analyses that must be performed (say a many-station, many-variable trend study) to work without detailed case-by-case checking of assumptions, then nonparametric procedures are ideal. They are always nearly as powerful as regression, and the failure to edit out or correctly transform a small percentage of outlying data will not have a substantial effect on the results.

Example 1

Appendix C10 lists phosphorus loads and streamflow during 1974-1985 on the Illinois River at Marseilles, IL. The Mann-Kendall and regression lines are plotted along with the data in figure 12.4. Both lines have slopes not significantly different from zero at $\alpha = 0.05$. The large load at the beginning of the record and non-normality of data around the regression line are the likely reasons the regression is considerably less significant. Improvements to the model are discussed in the next sections.

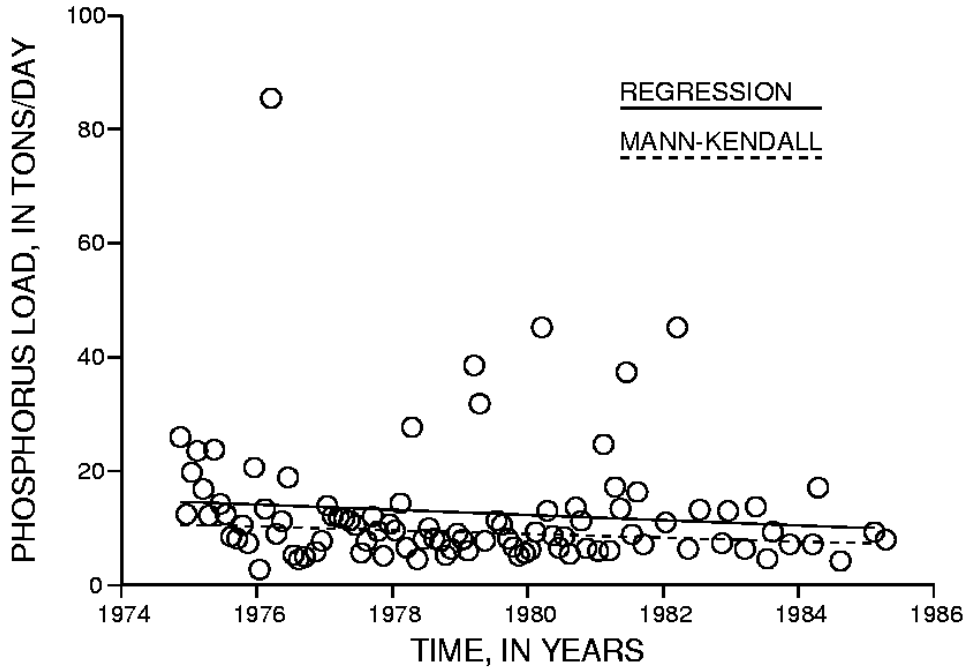


Figure 12.4 Mann-Kendall and regression trend lines (data in Appendix C10).

Regression:	Load =	16.8 – 0.46•time	t = –1.09	p = 0.28
Mann-Kendall:	Load =	12.2 – 0.33•time	tau = –0.12	p = 0.08.

12.3 Accounting for Exogenous Variables

Variables other than time trend often have considerable influence on the response variable Y. These "exogenous" variables are usually natural, random phenomena such as rainfall, temperature or streamflow. By removing the variation in Y caused by these variables, the background variability or "noise" is reduced so that any trend "signal" present can be seen. The ability (power) of a trend test to discern changes in Y with T is then increased. The removal process involves modelling, and thus explaining, the effect of exogenous variables with regression or LOWESS (for computation of LOWESS, see Chapter 10). This is the rationale for using the methods in the right-hand column of table 12.2.

For example, figure 12.5a presents a test for trend in dissolved solids at the James River in South Dakota. No adjustment for discharge was employed. The p-value for the test equals 0.47, so no trend is able to be seen. The Theil estimate of slope is plotted, showing the line

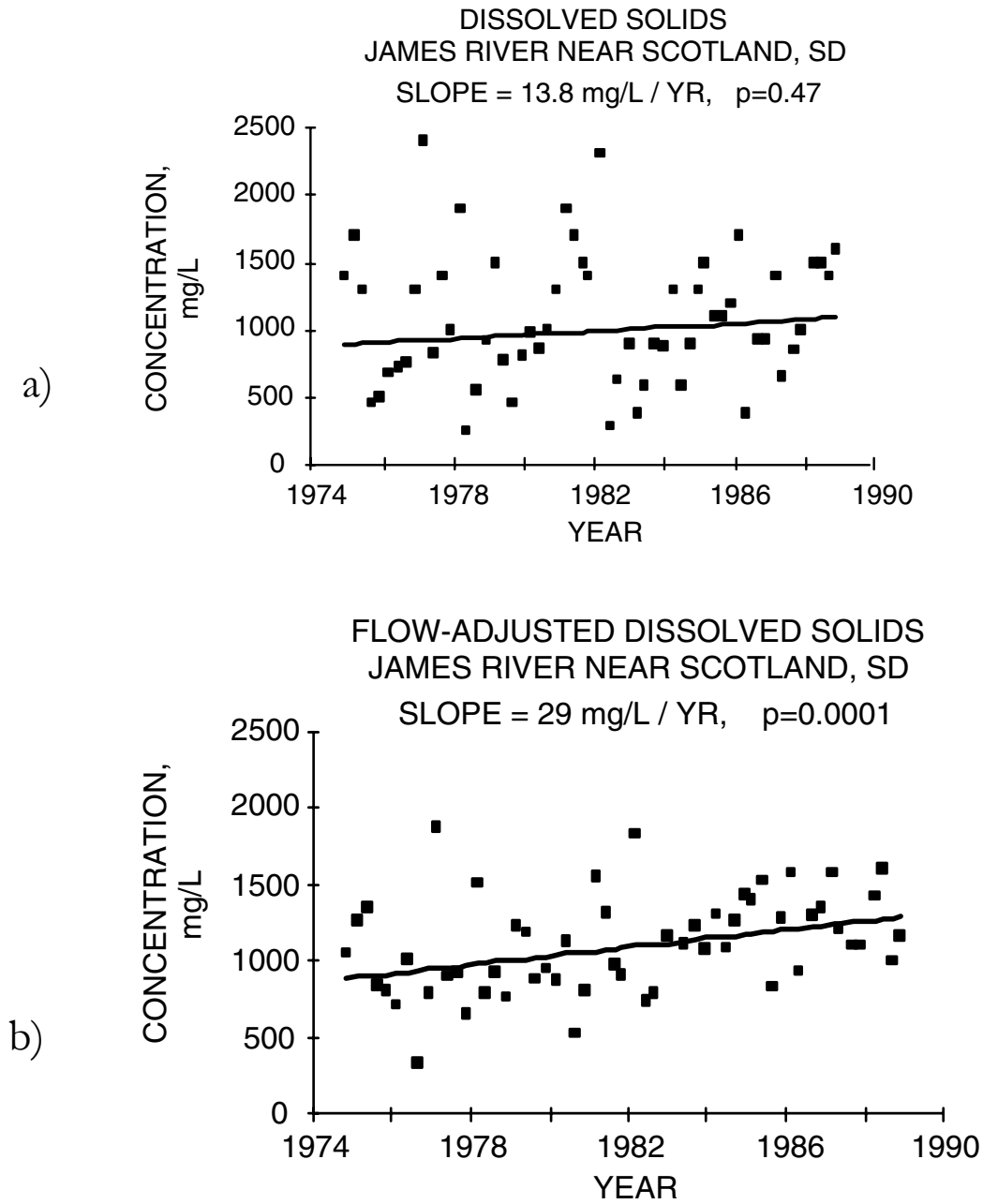


Figure 12.5 Trend tests a) before adjustment for flow. b) after adjustment for flow. (from Hirsch et al., 1991)

to be surrounded by a lot of data scatter. In figure 12.5b, the same data are plotted after using regression to remove the variation due to discharge. Note that the amount of scatter has

decreased. The same test for trend now has a p-value of 0.0001; for a given magnitude of flow, dissolved solids are increasing over time.

When removing the effect of one or more exogenous variables X, the probability distribution of the Xs is assumed to be unchanged over the period of record. Consider a regression of Y versus X (figures 12.6a and 6b). The residuals R from the regression describe the values for the Y variable "adjusted for" exogenous variables (figure 12.6c). In other words, the effect of other variables is removed by using residuals -- residuals express the variation in Y over and above that due to the variation caused by changes in the exogenous variables. A trend in R implies a trend in the relationship between X and Y (figure 12.6d). This in turn implies a trend in the distribution of Y over time while accounting for X. However, if the probability distribution of the Xs has changed over the period of record, a trend in the residuals may not necessarily be due to a trend in Y.

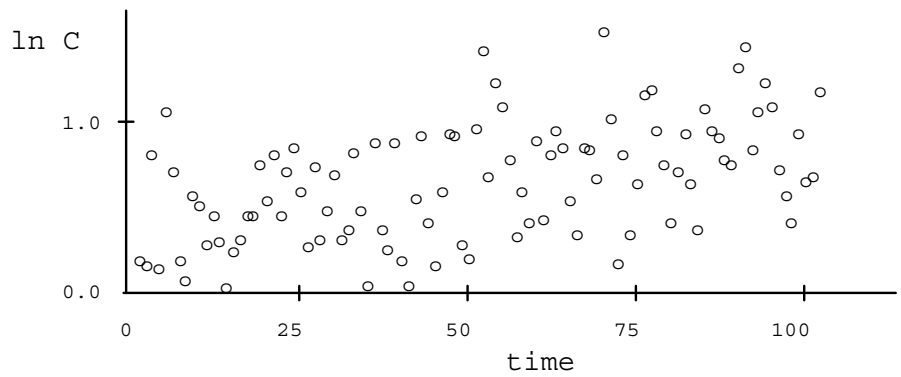


Figure 12.6a. Log of concentration vs. time. Trend is somewhat difficult to see.

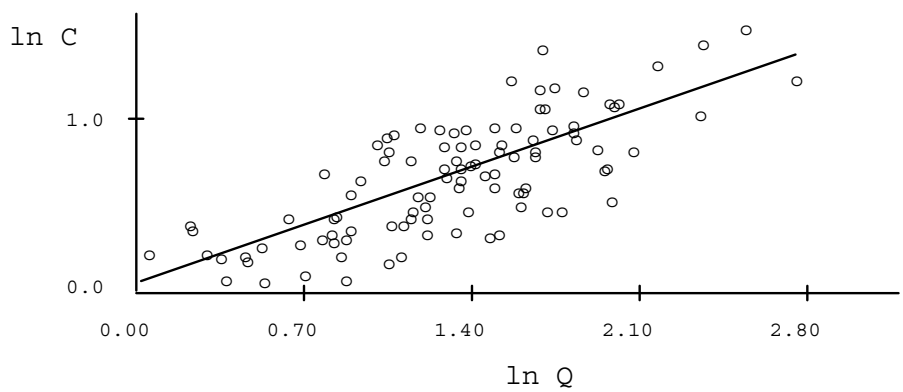


Figure 12.6b. Ln of concentration vs. exogenous variable: ln of streamflow (Q).
Strong linear relation shown by regression line.
Expect higher concentrations at higher flows.

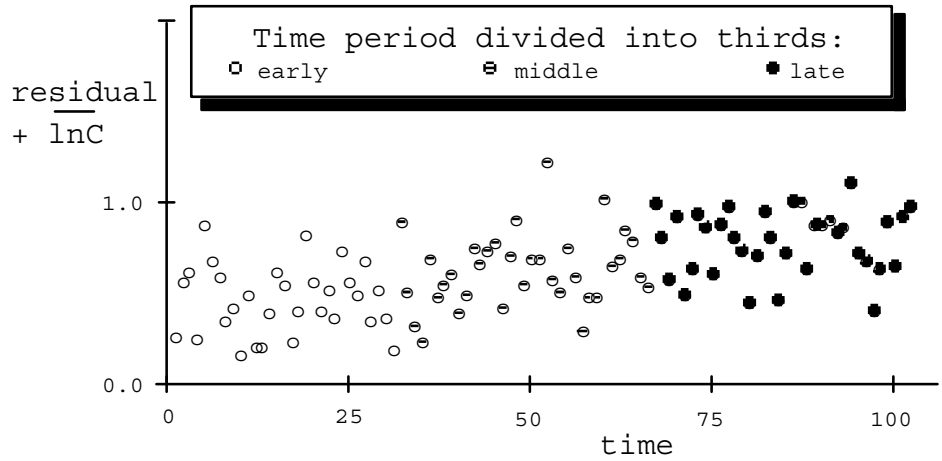


Figure 12.6c. Residuals from 12.6b regression over time. Trend much easier to detect than in 12.6a, as effect of Q has been removed by using residuals.

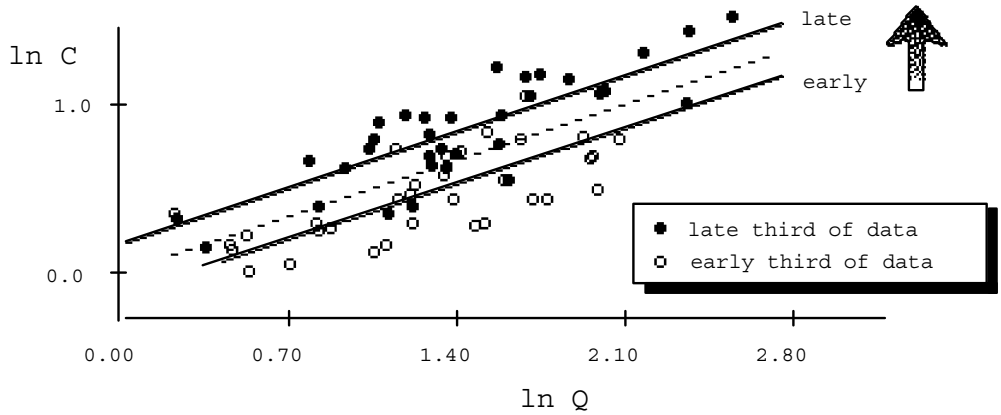


Figure 12.6d. Trend in fig. 6c can also be seen as an increase in the lnC vs lnQ relationship over time. For a given value of Q, the value for C increases over time.

What kind of variable is appropriate to select as an exogenous variable? It should be a measure of the driving force behind the process of interest, but must be free of changes in human manipulation. Thus a streamflow record that spans a major reservoir project, new diversion, or new operating policy on an existing system would be unacceptable, due to human alteration of the probability distribution of X during the period of interest. A streamflow record which reflects some human influence is acceptable, provided that the human influence is consistent over the period of record. Where human influence on streamflow records makes them unacceptable as X variables, two major alternatives exist. The first is to use flow at a nearby unaffected station which could be expected to be correlated with natural flow at the site of interest. The other alternative is to use weather-related data: rainfall over some antecedent period, or model-generated streamflows resulting from a deterministic watershed model that is driven by historical weather data.

Where Y is streamflow concentration, a great deal of the variance in Y is usually a function of river discharge. This comes about as a result of two different kinds of physical phenomena. One is dilution: a solute may be delivered to the stream at a reasonably constant rate (due to a point source or ground-water discharge to the stream) as discharge changes over time. The result of this situation is a decrease in concentration with increasing flow (see figure 12.7). This is typically seen in most of the major dissolved constituents (the major ions). The other process is wash-off: a solute, sediment, or a constituent attached to sediment can be delivered to the stream primarily from overland flow from paved areas or cultivated fields, or from streambank erosion. In these cases, concentrations as well as fluxes tend to rise with increasing discharge (see fig. 12.8). Some constituents can exhibit combinations of both of these kinds of behavior. One example is total phosphorus. A portion of the phosphorus may come from point sources such as sewage treatment plants (dilution effect), but another portion may be derived from surface wash-off and be attached to sediment particles (see fig. 12.9). The resulting pattern is an initial dilution, followed by a stronger increase with flow due to washoff.

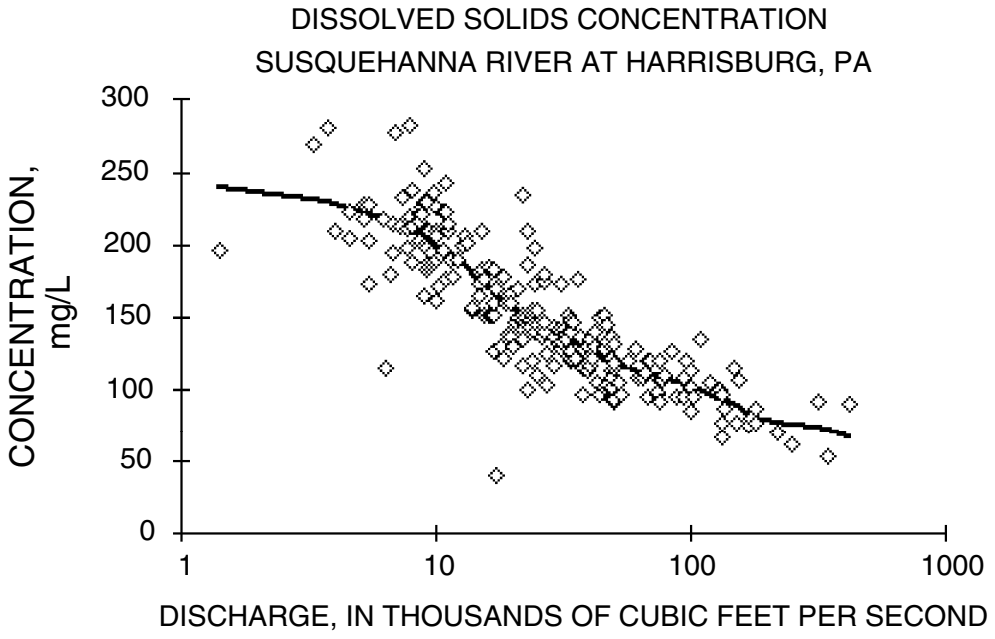


Figure 12.7 Dilution of dissolved solids with discharge (from Hirsch et al., 1991).

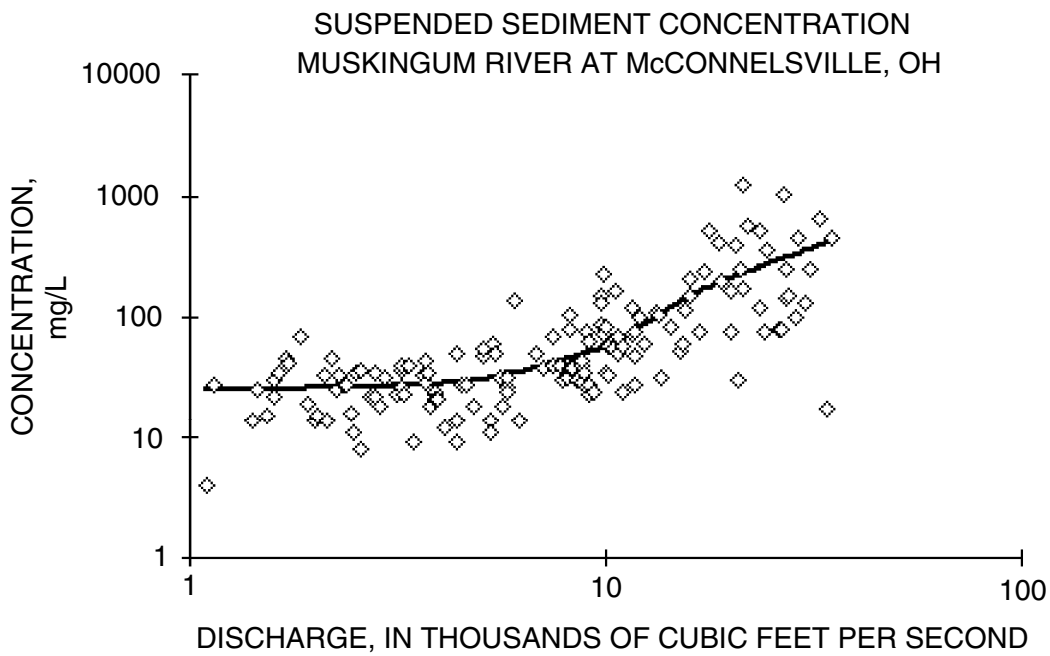


Figure 12.8 Washoff of suspended sediment with discharge (from Hirsch et al., 1991).

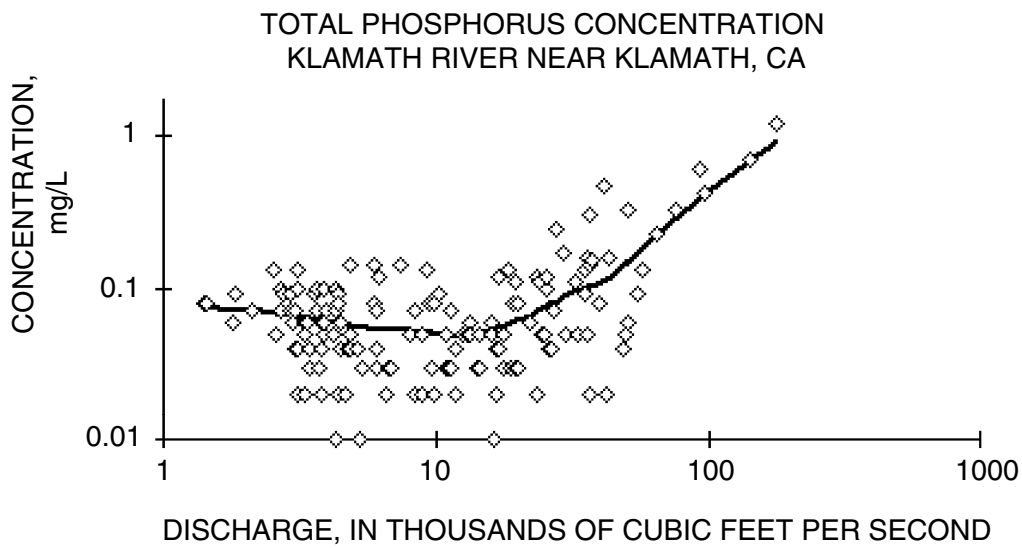


Figure 12.9 Dilution and subsequent washoff of total phosphorus as discharge increases (from Hirsch et al., 1991).

12.3.1 Nonparametric Approach

The smoothing technique LOWESS (LOcally WEighted Scatterplot Smooth) describes the relationship between Y and X without assuming linearity or normality of residuals. It is a robust description of the data pattern. Numerous smooth patterns result whose form changes as the

smoothing coefficient is altered. The LOWESS pattern chosen should be smooth enough that it doesn't have several local minima and maxima, but not so smooth as to eliminate true changes in slope. Given the LOWESS fitted value $\hat{Y}$ the residuals R are computed as

$$R = Y - \hat{Y} \ .$$

Then the Kendall S statistic is computed from the R,T data pairs, and tested to see if it is significantly different from zero. The test for S is the test for trend.

12.3.2 Mixed Approach: Mann-Kendall on Regression Residuals

To remove the effect of X on Y prior to performing the Mann-Kendall test, a linear regression could be computed between Y and one or more Xs:

$$Y = \beta_0 + \beta_1 \bullet X + \varepsilon$$

Unlike LOWESS, the adequacy of the regression model (is β_1 significant, should X be transformed due to lack of linearity or constant variance?) must be checked. When all is OK, the residuals R are computed as observed minus predicted values:

$$R = Y - b_0 - b_1 \bullet X \ .$$

Then the Kendall S statistic is computed from the R,T data pairs, and tested to see if it is significantly different from zero.

The Mann-Kendall test on residuals is a hybrid procedure -- parametric removal of effects of the exogenous variables, followed by a nonparametric test for trend. Care must be taken to insure that the model of Y versus X is reasonable (residuals should have no extreme outliers, Y is linear with X, etc.). The fully nonparametric alternative using LOWESS given in 12.3.1 avoids the necessity for close checking of assumptions.

Alley (1988) showed that this two-stage procedure resulted in lower power than an alternative which is analogous to the partial plots of Chapter 9. His "adjusted variable Kendall test" performs the second stage as a Mann-Kendall test of R versus T* rather than R versus T, where T* are the residuals from a regression of T versus X:

$$T = b_0 + b_1 \bullet X + T^*$$

In this way the effect of a drift in X over time is removed, so that the R versus T* relationship is totally free of the influence of X. This test is a Mann-Kendall test on the partial plot of Y versus T, having removed the effect of all other exogenous variable(s) X from both Y and T by regression.

12.3.3 Parametric Approach

Consider the multiple regression of Y versus time T and one or more Xs:

$$Y = \beta_0 + \beta_1 \bullet T + \beta_2 \bullet X + \varepsilon.$$

The null hypothesis for the trend test is that $\beta_1 = 0$. Therefore the t-statistic for β_1 tests for trend. This test simultaneously compensates for the effects of exogenous variables by including them in the model. No two-stage process is necessary. The model must be checked for

adequacy – for the correct form of relationship (linear in the Xs and T), normality of residuals, and that b_2 is significantly different from zero. If b_1 is significantly different from zero (based on the t-statistic) then the null hypothesis of no trend is rejected, and we conclude that there is a linear trend in Y over T.

12.3.4 Comparison of Approaches

In general, the power and efficiency of any procedure for detecting and estimating the magnitude of trends will be aided if the variance of the data can be decreased (figure 12.5). This can be done by removing discharge effects either simultaneously or in stages. Simultaneous modelling of trend and discharge has a small but distinct advantage over the equivalent stagewise method (Alley, 1988). Thus parametric multiple regression has more power than a stagewise regression. The adjusted Kendall test has a similar advantage over the Mann-Kendall test on residuals R versus unadjusted T. We presume that a Mann-Kendall test of R on T* where both are computed using LOWESS (Y on X and T on X) would have similar advantages over the unadjusted method in section 12.3.1, though no data exists on this to date.

More important is whether the adjustment process should be conducted using a parametric or nonparametric method. The choice between regression and LOWESS should be based on the quality of the regression fit. LOWESS and linear regression fits of phosphorus concentration and stream discharge are compared for Klamath River in Figure 12.10. LOWESS would be a sensible alternative here due to the nonlinearity of the relationship. In studies where many data sets are being analyzed, and individualized checking of multiple models is impractical, LOWESS is the method of choice. It is also valuable when transformation of Y to achieve normality is not desirable. Where detailed model checking is practical and where high-quality parametric models can be constructed, multiple regression provides a one-step process with maximum efficiency. It and the adjusted Kendall method should be used over stagewise procedures.

All methods incorporating exogenous X variables discussed thus far assume that the change in the X,Y relationship over time is a parallel shift -- a change in intercept, no change in slope (see figure 12.6d). Changes in both (a rotation) are certainly possible. However it will not be possible to classify all such changes as uptrends or downtrends. For example, if the X,Y relationship pivots counterclockwise over time, then for high X there is an uptrend in Y and for low X there is a downtrend in Y. There is no simple way to generalize the Mann-Kendall test on residuals to identify such situations.

However, regression could be used as follows:

$$Y = \beta_0 + \beta_1 \bullet X + \beta_2 \bullet T + \beta_3 \bullet X \bullet T + \varepsilon$$

The X•T term is an interaction term describing the rotation. One could compare this model to the "no trend" model (one with no T or X•T terms) using an F test. It can also be compared to

the simple trend model (one with an X and a T term but no X•T term) using a partial F test. The result will be selection of one of three outcomes: no trend, trend in the intercept, or trend in slope and intercept (rotation).

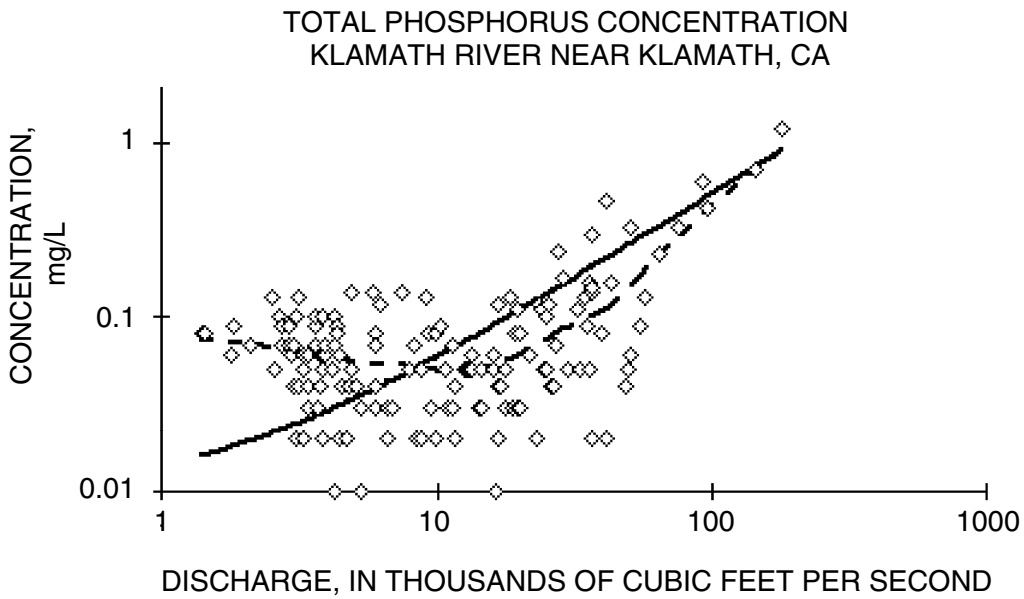


Figure 12.10 Comparison of LOWESS (dashed line) and linear regression (solid line) fits of concentration to stream discharge. From Hirsch, et al. (1991).

12.4 Dealing With Seasonality

There are many instances where changes between different seasons of the year are a major source of variation in the Y variable. As with other exogenous effects, seasonal variation must be compensated for or "removed" in order to better discern the trend in Y over time. If not, little power may be available to detect trends which are truly present. We may also be interested in modeling the seasonality to allow different predictions of Y for differing seasons.

Most concentrations in surface waters show strong seasonal patterns. Streamflow itself almost always varies greatly between seasons. This arises from seasonal variations in precipitation volume, and in temperature which in turn affects precipitation type (rain versus snow) and the rate of evapotranspiration. Some of the observed seasonal variation in water quality may be explained by accounting for this seasonal variation in discharge. However, seasonality often remains even after discharge effects have been removed (Hirsch et al. 1982). Possible additional causes of seasonal patterns include biological activity, both natural and managed activities such as agriculture. For example, nutrient concentrations vary with seasonal application of fertilizers and the natural pattern of uptake and release by plants. Other effects are due to different

sources of water dominant at different times of the year, such as snow melt versus intense rainfall. Seasonal rise and fall of ground water can also influence water quality. A given discharge in one season may derive mostly from ground water while the same discharge during the another season may result from surface runoff or quick flow through shallow soil horizons. The chemistry and sediment content of these sources may be quite different.

Techniques for dealing with seasonality fall into three major categories (table 12.3). One is fully nonparametric, one is a mixed procedure, and the last is fully parametric. In the first two procedures it is necessary to define a "season". In general, seasons should be just long enough so that there is some data available for most of the seasons in most of the years of record. For example, if the data are primarily collected at a monthly frequency, the seasons should be defined to be the 12 months. If the data are collected quarterly then there should be 4 seasons, etc. Tests for trend listed in table 12.2 have analogs which deal with seasonality. These are presented in table 12.3.

	Not Adjusted for X	**Adjusted for X**
Nonparametric	Seasonal Kendall test for trend on Y (Method I)	Seasonal Kendall trend test on residuals from LOWESS of Y on X (Method I)
Mixed	Regression of deseasonalized Y on T (Method II)	Seasonal Kendall trend test on residuals from regression of Y on X (Method I)
Parametric	Regression of Y on T and seasonal terms (Method III)	Regression of Y on X, T, and seasonal terms (Method III)

Table 12.3 Methods for dealing with seasonal patterns in trend testing

12.4.1 The Seasonal Kendall Test

The seasonal Kendall test (Hirsch et al., 1982) accounts for seasonality by computing the Mann-Kendall test on each of m seasons separately, and then combining the results. So for monthly "seasons", January data are compared only with January, February only with February, etc. No comparisons are made across season boundaries. Kendall's S statistic S_i for each season are summed to form the overall statistic S_k.

$$S_k = \sum_{i=1}^{m} S_i \qquad [12.1]$$

When the product of number of seasons and number of years is more than about 25, the distribution of S_k can be approximated quite well by a normal distribution with expectation equal to the sum of the expectations (zero) of the individual S_i under the null hypothesis, and variance equal to the sum of their variances. S_k is standardized (eq. 12.2) by subtracting its expectation $\mu_k = 0$ and dividing by its standard deviation σ_{Sk}. The result is evaluated against a table of the standard normal distribution.

$$Z_{Sk} = \begin{cases} \dfrac{S_k - 1}{\sigma_{Sk}} & if\ S_k > 0 \\ 0 & if\ S_k = 0 \\ \dfrac{S_k + 1}{\sigma_{Sk}} & if\ S_k < 0 \end{cases} \qquad [12.2]$$

where $\mu_{Sk} = 0$,

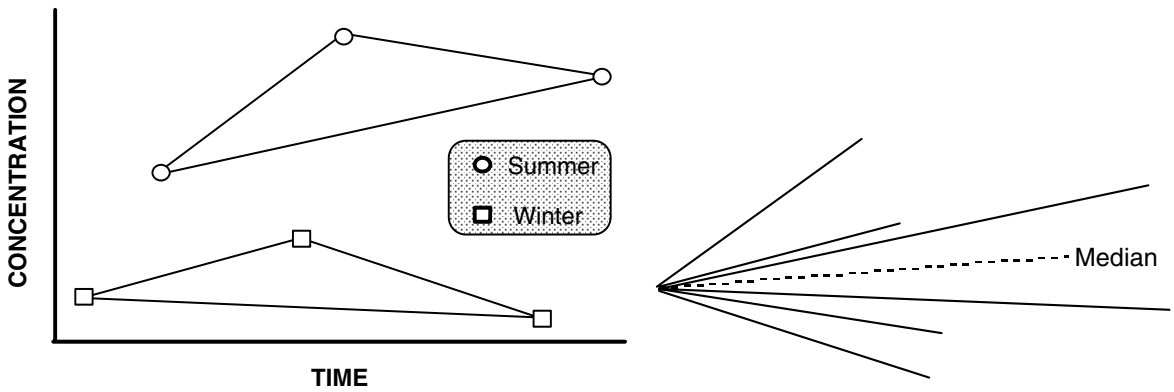
$$\sigma_{Sk} = \sqrt{\sum_{i=1}^{m} (n_i/18)\bullet(n_i-1)\bullet(2n_i+5)}\ , \quad \text{and}$$

n_i = number of data in the ith season.

The null hypothesis is rejected at significance level α if $|Z_{Sk}| > Z_{crit}$ where Z_{crit} is the value of the standard normal distribution with a probability of exceedance of $\alpha/2$. When some of the Y and/or T values are tied the formula for σ_{Sk} must be modified, as discussed in Chapter 8. The significance test must also be modified for serial correlation between the seasonal test statistics (see Hirsch and Slack, 1984).

If there is variation in sampling frequency during the years of interest, the data set used in the trend test may need to be modified. If variations in sampling frequency are random (for example if there are a few instances where no value exists for some season of some year, and a few instances when two or three samples are available for some season of some year) then the data can be collapsed to a single value for each season of each year by taking the median of the available data in that season of that year. If, however, there is a systematic trend in sampling frequency (monthly for 7 years followed by quarterly for 5 years) then the following type of approach is necessary. Define the seasons on the basis of the lowest sampling frequency. For that part of the record with a higher frequency define the value for the season as the observation taken closest to the midpoint of the season. The reason for not using the median value in this case is that it will induce a trend in variance, which will invalidate the null distribution of the test statistic.

An estimate of the trend slope for Y over time T can be computed as the median of all slopes between data pairs within the same season (figure 12.11). Therefore no cross-season slopes contribute to the overall estimate of the Seasonal Kendall trend slope.



A. B.

Figure 12.11 A. All pairwise slopes used to estimate the Seasonal Kendall trend slope (two seasons -- compare with figure 10.1).
B. Slopes rearranged to meet at a common origin

To accommodate and model the effects of exogenous variables, directly follow the methods of section 12.3 until the final step. Then apply the Seasonal Kendall rather than Mann-Kendall test on residuals from a LOWESS of Y versus X and T versus X
(R versus T*).

12.4.2 Mixture Methods

The seasonal Kendall test can be applied to residuals from a regression of Y versus X, rather than LOWESS. Keep in mind the discussion in the previous section of using adjusted variables T* rather than T. Regression would be used only when the relationships exhibit adherence to the appropriate assumptions.

A second type of mixed procedure involves deseasonalizing the data by subtracting seasonal medians from all data within the season, and then regressing these deseasonalized data against time. One advantage of this procedure is that it produces a description of the pattern of the seasonality (in the form of the set of seasonal medians). However, this method has generally lower power to detect trend than other methods, and is not prefered over the other alternatives. Subtracting seasonal means would be equivalent to using dummy variables for m−1 seasons in a fully parametric regression. Either use up m−1 degrees of freedom in computing the seasonal statistics, a disadvantage which can be avoided by using the methods of the next section.

12.4.3 Multiple Regression With Periodic Functions

The third option is to use periodic functions to describe seasonal variation. The simplest case, one that is sufficient for most purposes, is:

$$Y = \beta_0 + \beta_1 \bullet \sin(2\pi T) + \beta_2 \bullet \cos(2\pi T) + \beta_3 \bullet T + \text{other terms} + \varepsilon \qquad [12.3]$$

where "other terms" are exogenous explanatory variables such as flow, rainfall, or level of some human activity (e.g. waste discharge, basin population, production). They may be continuous, or binary "dummy" variables as in analysis of covariance. The trend test is conducted by determining if the slope coefficient on T (β_3) is significantly different from zero. Other terms in the equation should be significant and appropriately modeled. The residuals ε must be approximately normal.

Time is commonly but not always expressed in units of years. Table 12.4 lists values for $2\pi T$ for three common time units: years, months and day of the year.

The expression	$2\pi T$	$= 6.2832 \bullet t$	when t is expressed in years.
		$= 0.5236 \bullet m$	when m is expressed in months.
		$= 0.0172 \bullet d$	when d is expressed in day of year.

Table 12.4 Three values for $2\pi T$ useful in regression tests for trend

To more meaningfully interpret the sine and cosine terms, they can be re-expressed as the amplitude A of the cycle (half the distance from peak to trough) and the day of the year Dp at which the peak occurs:

$$\beta_1 \bullet \sin(2\pi t) + \beta_2 \bullet \cos(2\pi t) = A \sin\left[2\pi(t + t_0)\right] \qquad [12.4]$$

where $A = \sqrt{\beta_1^2 + \beta_2^2}$ [12.5]

The phase shift $t_0 = \tan^{-1}(\beta_2 / \beta_1)$,

$t_0' = t_0 \pm 2\pi$ if necessary to get t_0 within the interval $0 < t_0 < 2\pi = 6.2832$

and $Dp = 58.019 \bullet (1.5708 - t_0')$ [12.6]

Example 2: Determining peak day and amplitude

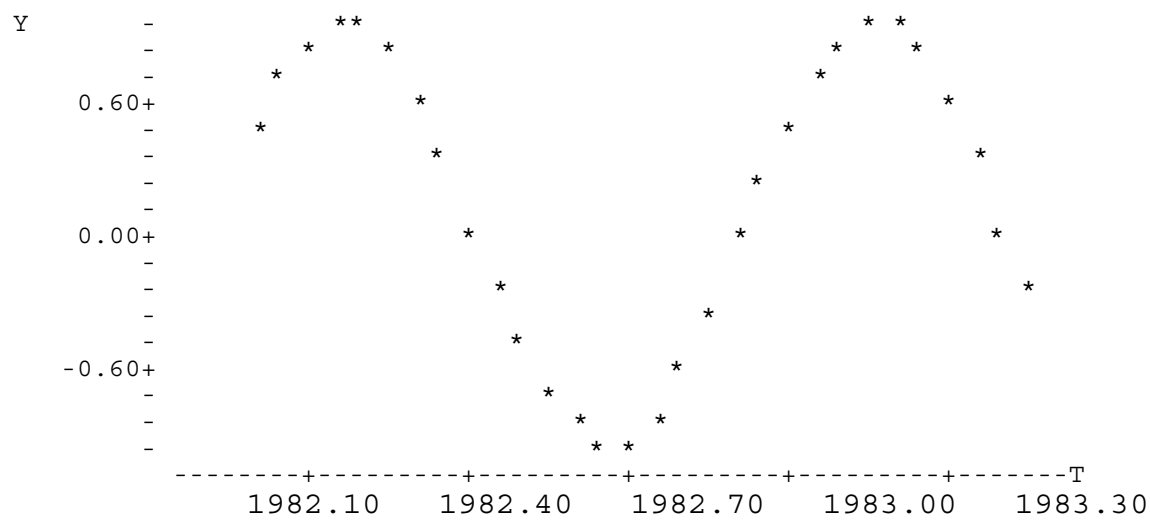


Figure 12.12 Data showing seasonal (sine and cosine) pattern.

The data in figure 12.12 were generated using coefficients $b_1 = 0.8$ and $b_2 = 0.5$. From equations 12.4 through 12.6, the amplitude A = 0.94, $t_0 = 0.559$, $t_0' = 0.559$, and so the peak day = 59 (February 28).

After including sine and cosine terms in a multiple regression to account for seasonality, the residuals may still show a seasonal pattern in boxplots by season, or in a Kruskal-Wallis test by season. If this occurs, additional periodic functions with periods of 1/2 or 1/3 or other fractions of a year (multiple cycles per year) may be used to remove additional seasonality. Credible explanations for why such cycles might occur are always helpful. For example, the following equation may be appropriate:

$$Y = \beta_0 + \beta_1 \bullet \sin(2\pi t) + \beta_2 \bullet \cos(2\pi t) + \beta_3 \bullet \sin(4\pi t) + \beta_4 \bullet \cos(4\pi t) + \text{other terms} + \varepsilon$$

One way to determine how many terms to use is to add them, two at a time, to the regression and at each step do an F test for the significance of the new pair of terms. As a result one may, very legitimately, settle on a model in which the t-statistics for one of a pair of coefficients is not significant, but as a pair they are significant. Leaving out just the sine or just the cosine is not a sensible thing to do, because it forces the periodic term to have a completely arbitrary phase shift, rather than one determined by the data.

12.4.4 Comparison of Methods

The Mann-Kendall and mixed approaches have the disadvantages of only being applicable to univariate data (either original units or residuals from a previous analysis) and are not amenable to simultaneous analysis of multiple sources of variation. They take at least two steps to

compute. Multiple regression allows many variables to be considered easily and simultaneously by a single model.

Mann-Kendall has the usual advantage of nonparametrics: robustness against departures from normality. The mixed method is perhaps the least robust because the individual seasonal data sets can be quite small and the estimated seasonal medians can follow an irregular pattern. In general this method has far more parameters than either of the other two methods and fails to take advantage of the idea that geophysical processes have some degree of smoothness in the annual cycle. That is: it is unlikely that April will be very different from May, even though the sample statistics may suggest that this is so.

Regression with periodic functions takes advantage of this notion of smoothness and thereby involves very few parameters. However, the functional form (sine and cosine terms) can become a "straight jacket". Perhaps the annual cycles really do have abrupt breaks associated with freezing and thawing, or the growing season. Regression can always use binary variables as abrupt definitions of season (G=1 for "growing season", G=0 otherwise). Observations can be assigned to a season based on conditions which may vary in date from year to year, and not just based on the date itself. Regression could also be modified to accept other periodic functions, perhaps ones that are much more squared off. To do this demands a good physically-based definition of the timing of the influential factors, however.

All three methods provide a description of the seasonal pattern. Regression and mixed methods automatically produce seasonal summary statistics. However, there is no difficulty in providing a measure of seasonality consistent with Mann-Kendall by computing seasonal medians of the data after trend effects have been removed.

12.4.5 Presenting Seasonal Effects

There are many ways of characterizing the seasonality of a data set (table 12.5). Any of them can be applied to the raw data or to residuals from a LOWESS or regression that removes the effects of some exogenous variable. In general, graphical techniques will be more interpretable than tabular, although the detail of tables may sometimes be needed.

	GRAPHICAL METHODS	TABULAR METHODS
BEST	Boxplots by season, or LOWESS of data versus time of year	List the amplitude and peak day of the cycle
NEXT BEST		List of seasonal medians and seasonal interquartile ranges, or list of distribution percentage points by season
WORST	Plot of seasonal means with standard deviation or standard error bars around them	List of seasonal means, standard deviations, or standard errors

Table 12.5 Rating of methods for dealing with seasonality

12.4.6 Differences Between Seasonal Patterns

The approaches described above all assume a single pattern of trend across all seasons. This may be a gross over-simplification and can fail to reveal large differences in behavior between different seasons. It is entirely possible that the Y variable exhibits a strong trend in its summer values and no trend in the other seasons. Even worse, it could be that spring and summer have strong up-trends and fall and winter have strong down-trends, cancelling each other out and resulting in an overall seasonal Kendall test statistic stating no trend. Another situation might arise where the X-Y relationship (e.g. rainfall-runoff, flow-concentration) has a substantially different slope and intercept for different seasons.

No overall test statistic will provide any clue of these differences. This is not to suggest they are not useful. Many times we desire a single number to characterize what is happening in a data set. Particularly when dealing with several data sets (multiple stations and/or multiple variables), breaking the problem down into 4 seasons or 12 months simply swamps us with more results than can be absorbed. Also, if the various seasons do show a consistent pattern of behavior there is great strength in looking at them in one analysis. For example in a seasonal Kendall analysis each month viewed by itself might show a positive S value, none of which is significant, but the overall Seasonal Kendall test could be highly significant. Yet in detailed examinations of individual stations it is often useful to perform and present the full, within-season analysis on each season. Figure 12.13 is a good approach to graphically presenting the results of such multi-season analyses.

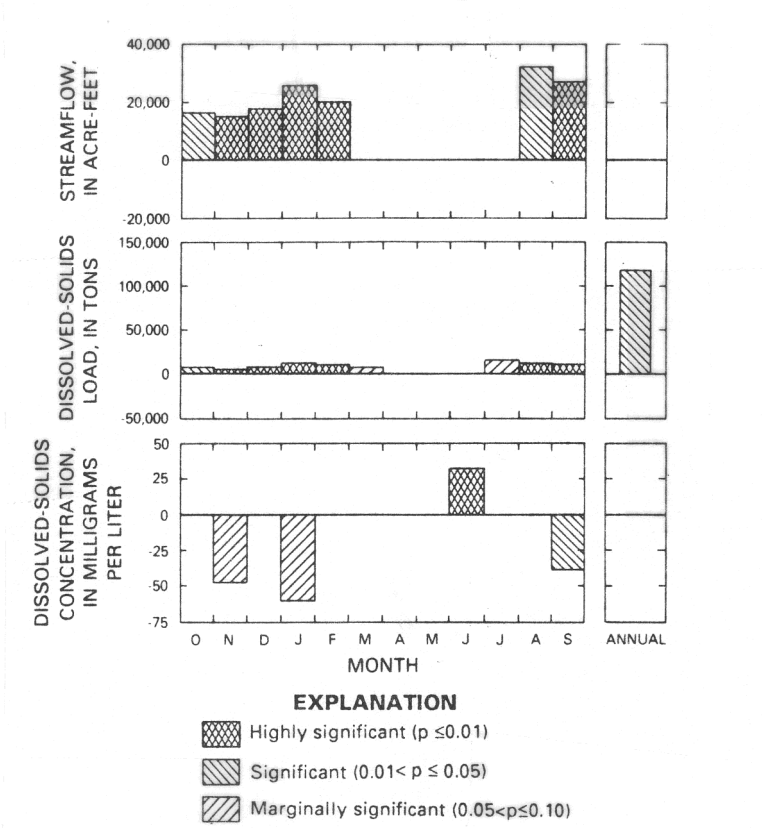


Figure 12.13 Illustration of seasonal and annual step trends on the Green River (from Liebermann and others, 1989)

In the approaches using Method I above, one can also examine "contrasts" between the different seasonal statistics. This provides a single statistic which indicates whether the seasons are behaving in a similar fashion (homogeneous) or behaving differently from each other (heterogeneous). The test for homogeneity is described by van Belle and Hughes (1984).

For each season i (i=1,2,...m) compute $Z_i = S_i / \sqrt{Var(S_i)}$. Sum these to compute the "total" chi-square statistic, then compute "trend" and "homogeneous" chi-squares:

$$\chi^2_{(total)} = \sum_{i=1}^{m} Z_i^2 \qquad [12.7]$$

$$\chi^2_{(trend)} = m \bullet \bar{Z}^2 \qquad \text{where } \bar{Z} = \frac{\sum_{i=1}^{m} Z_i}{m} \qquad [12.8]$$

$$\chi^2_{(homogeneous)} = \chi^2_{(total)} - \chi^2_{(trend)} \qquad [12.9]$$

The null hypothesis that the seasons are homogeneous with respect to trend ($\tau_1 = \tau_2 = \ldots = \tau_m$) is tested by comparing $\chi^2_{(homogeneous)}$ to tables of the chi-square distribution with m−1 degrees of freedom. If it exceeds the critical value for the pre-selected α, reject the null hypothesis and conclude that different seasons exhibit different trends.

12.5 Use of Transformations in Trend Studies

Water resources data commonly exhibit substantial departures from a normal distribution. Surface-water concentration, load, and flow data are often positively skewed, with many observations lying close to a lower bound of zero and a few observations one or more orders of magnitude above the lower values. If only a test for trend is of interest, then the decision to make some monotonic transformation of the data (to render them more nearly normal) is of no consequence provided that a nonparametric test is used. Nonparametric trend tests are invariant to monotonic power transformations (such as the logarithm or square root). In terms of significance levels the test results will be identical whether the test was applied to the raw data or the transformed data.

The decision to transform data is, however, highly important in terms of any of the procedures for removing the effects of exogenous variables (X), for computing significance levels of a parametric test (figure 12.14), and for computing and expressing slope estimates. Trends which are nonlinear (say exponential or quadratic) will be poorly described by a linear slope coefficient, whether from regression or a nonparametric method. It is quite possible that negative predictions may result for some values of time or X. By transforming the data so that the trend is linear, a Mann-Kendall or regression slope can later be re-expressed back into original units. The resulting nonlinear trend will better fit the data than the linear expression, even though their nonparametric significance tests are identical. Thus, it may be appropriate to run analyses on transformed Y values, even if the analysis is a nonparametric one.

One way to ensure that the fitted trend line will not predict negative values is to take a log transformation of the data prior to trend analysis. The trend slope will then be expressed in log units. A linear trend in log units translates to an exponential trend in original units, which can then be re-expressed in percent per year to make the trend easier to interpret. If b_1 is the estimated slope of a linear trend in natural log units then the percentage change from the beginning of any year to the end of that year will be $(e^{b1} - 1)\bullet 100$. If slopes in original units are preferred, then instead of multiplying by 100, multiply by some measure of central tendency in the data (mean or median) to express the slope or step-trend in original units.

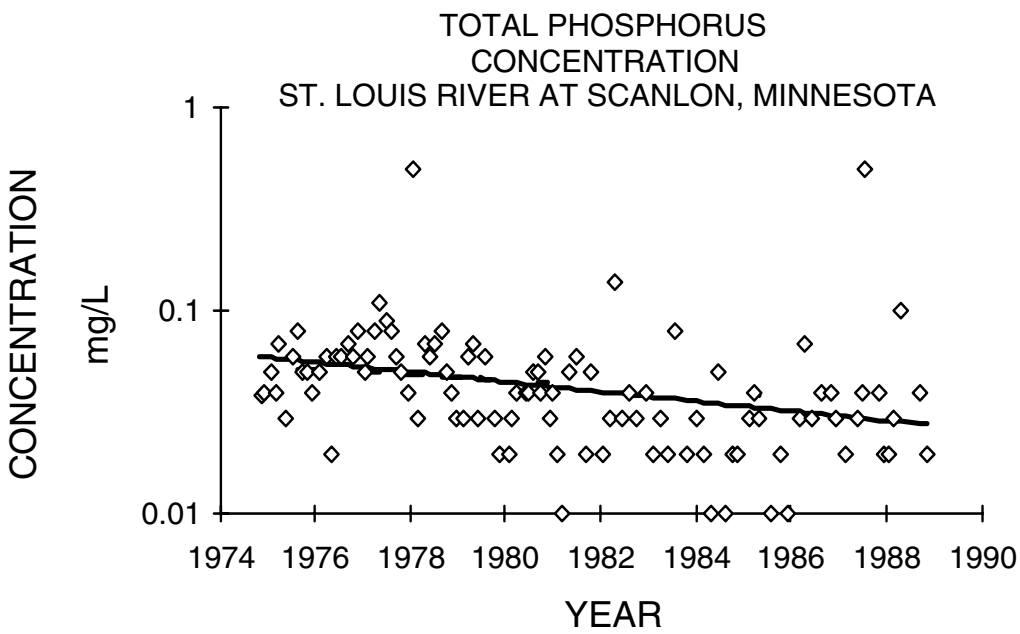


Figure 12.14 No trend evident in concentration (p=0.432 for regression slope). After log transformation, there is a statistically significant decline (p=0.001). Regression on logs shown as solid line. (from Hirsch et al., 1991).

In general, more resistant and robust results can be obtained if log transformations are used for variables that typically have ranges of more than an order of magnitude. With variations this large, transformations should be used in conjunction with both parametric and nonparametric tests. However, in multiple record analyses the decision to transform should be made on the basis of the characteristics of the class of variables being studied, not on a case-by-case basis. Variables on which log transforms are typically helpful include: flood flows, low flows, monthly or annual flows in small river basins, concentrations of sediment, total concentration (suspended plus dissolved) for a constituent when the suspended fraction is substantial (for example phosphorus and some metals), concentrations or counts of organisms, concentrations of substances that arise from biological processes (such as chlorophyll), and downstream load for virtually any constituent.

Some argue that data should always be transformed to normality, and parametric procedures computed on the transformed data. Transformations to normality are not always possible, as some data are non-normal due not to skewness but to heavy tails of the distribution (see Schertz and Hirsch, 1985). The strongest argument for transformations are that regression methods allows simultaneous consideration of the effects of multiple exogenous variables along with temporal trend. Such simultaneous tests are more difficult with nonparametric techniques. Multivariate smoothing methods are available (Cleveland and Devlin, 1988) which at least allow removal of multiple exogenous effects in one step, but they are not implemented yet in any commercial software.

The weakest situations for parametric techniques on transformed data are for analyses of multiple data sets. The transformation appropriate to one data set may not be appropriate to another. If different transformations are used on different data sets then comparisons among results is difficult, if not impossible. Also, there is an element of subjectivity in the choice of transformation. The argument of the skeptic that: "You can always reach the conclusion you want if you manipulate the data enough" is not without merit. The credibility of results is enhanced if a single statistical method is used for all data sets in a study, and this is next to impossible with the several judgements of model adequacy required for parametric methods. Nonparametric procedures are therefore well suited to multi-record trend analysis studies. In analyses of individual records, use of transformations with parametric methods can be very appropriate.

12.6 Monotonic Trend versus Two Sample (Step) Trend

Study of long term changes in hydrologic variables can be carried out in either of two modes. Up to this point "monotonic trends" were discussed, gradual and continuing changes over time. The Mann-Kendall test and regression are the two basic tools used in this case. The other mode compares two non-overlapping sets of data, an "early" and "late" period of record. Changes between the periods are called "step trends", as values of Y step up or down from one time period to the next. Testing for differences between these two groups involves procedures similar or identical to those described in other chapters, including the rank-sum test, two-sample t-tests, and analysis of covariance. Each of them also can be modified to account for seasonality.

The basic parametric test for step trends is the two-sample t-test. See Chapter 5 for its computation. The magnitude of change is measured by the difference in sample means between the two periods. Helsel and Hirsch (1988) discuss the disadvantages of using a t-test for step trends on data which are non-normal -- loss of power, inability to incorporate data below the detection limit, and an inappropriate measure of the step trend size. The primary nonparametric alternative is the rank-sum test and associated Hodges-Lehmann (H-L) estimator of step-trend magnitude (see Chapter 5, and Hirsch, 1988). The H-L estimator is the median of all possible differences between data in the "before" and "after" periods. Table 12.6 summarizes the step-trend approaches not considering seasonality and 12.7 summarizes those which consider seasonality. The rank-sum test can be implemented in a seasonal manner just like the Mann-Kendall test, called the seasonal rank-sum test. It computes the rank-sum statistic separately for each season, sums the test statistics, their expectations and variances, and then evaluates the overall summed test statistic. The H-L estimator can be similarly modified by considering only data pairs within a given season.

	Not Adjusted for X	**Adjusted for X**
Nonparametric	Rank-sum test on Y	Rank-sum test on residuals from LOWESS of Y on X
Mixed	---	Rank-sum test on residuals from regression of Y on X
Parametric	Two sample t-test	Analysis of covariance of Y on X and group

Table 12.6 Step-trend tests (two-sample) which do not consider seasonality (note "group" refers to a dummy variable 0 for "before" and 1 for "after")

	Not Adjusted for X	**Adjusted for X**
Nonparametric	Seasonal rank-sum test on Y	Seasonal rank-sum test on residuals from LOWESS of Y on X
Mixed	Two-sample t test on deseasonalized Y	Seasonal rank-sum test on residuals from regression of Y on X
Parametric	Analysis of covariance of Y on seasonal terms and group	Analysis of covariance of Y on X, seasonal terms, and group

Table 12.7 Step-trend tests (two-sample) which do consider seasonality (note "group" refers to a dummy variable 0 for "before" and 1 for "after")

Step trend procedures should be used in two situations. The first is when the record (or records) being analyzed are naturally broken into two distinct time periods with a relatively long gap between them. There is no specific rule to determine how long the gap should be to make this the preferred procedure. If the length of the gap is more than about one-third the entire period of data collection, then the step trend procedure is probably best (see figure 12.15). In general, if the within-period trends are small in comparison to the between-period differences, then step-trend procedures should be used.

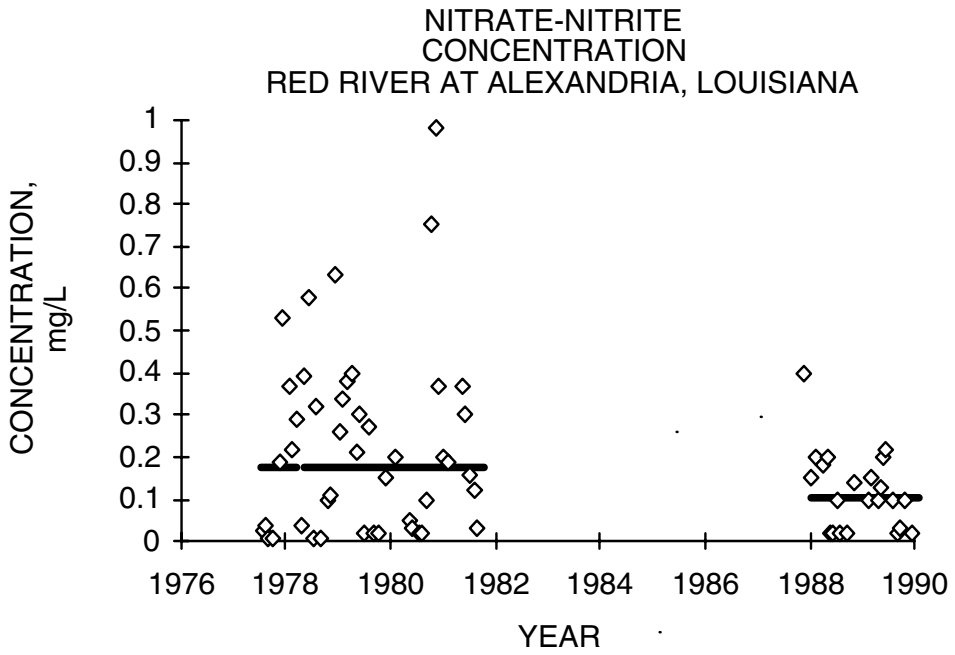


Figure 12.15 Significant (p=0.085) step trend as measured by rank-sum test. Solid lines are group medians. Monotonic trend test is not significant (p=0.167). Modified from Hirsch et al., 1991.

The second situation to test for step-trend is when a known event has occurred at a specific time during the record which is likely to have changed water quality. The record is first divided into "before" and "after" periods at the time of this known event. Example events are the completion of a dam or diversion, the introduction of a new source of contaminants, reduction in some contaminant due to completion of treatment plant improvements, or the closing of some facility (figure 12.16). It is imperative that the decision to use step-trend procedures not be based on examination of the data (i.e. the analyst notices an apparent step but had no prior hypothesis that it should have occurred), or on a computation of the time which maximizes the difference between periods. Such a prior investigation biases the significance level of the test, finding changes which are not really there. Step-trend procedures require a highly specific situation, and the decision to use them should be made prior to any examination of the data.

If there is no prior hypothesis of a time of change or if records from a variety of stations are being analyzed in a single study, monotonic trend procedures are most appropriate. In multiple record studies, even when some of the records have extensive but not identical gaps, the monotonic trend procedures are generally best because comparable periods of time are more easily examined among all the records. In fact, the frequent problem of multiple starting dates, ending dates, and gaps in a group of records presents a significant practical problem in trend analysis studies. In order to correctly interpret the data, records examined in a multiple station study must be concurrent. For example it is pointless to compare a 1975-1985 trend at one

station to a 1960-1980 trend at another. The difficulty arises in selecting a period which is long enough to be meaningful but does not exclude too many shorter records.

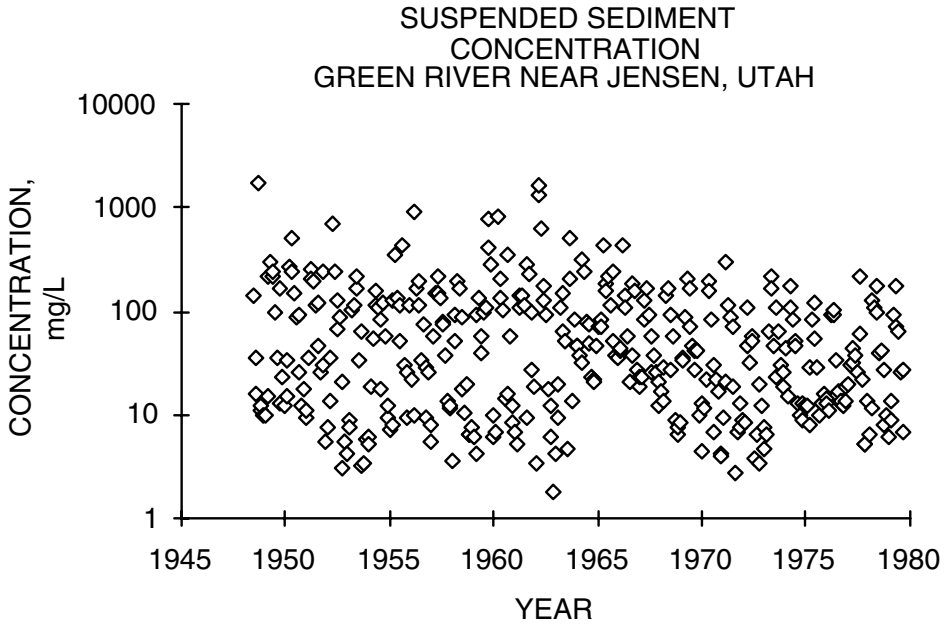


Figure 12.16 A weakly significant (p=0.105) reduction in suspended sediment after completion of the Flaming Gorge reservoir (located 93 miles upstream of the station) in late 1962 as measured by a rank-sum test. From Hirsch et al., 1991.

A further difficulty involves deciding just how complete a record must be to be included in the analysis. For example, if the study is for 1970-1985 and there is a record that runs from 1972 through 1985 it is probably prudent to include it in the study. A one- or two-year gap in the middle of the record should not disqualify a station from the analysis. More difficult are questions such as inclusion of a 1976-1984 record, or inclusion of a record that covers 1970-1975 and 1982-1985. One reasonably objective rule for deciding whether to include a record is:

1) divide the study period into thirds (three periods of equal length),
2) determine the coverage in each period (e.g. if the record is generally monthly, count the months for which there are data),
3) if any of the thirds has less than 20 percent of the total coverage then the record should not be included in the analysis. See Schertz (1990) for an application of these kinds of approaches.

12.7 Applicability of Trend Tests With Censored Data

Censored samples are records in which some of the data are known only to be "less than" or "greater than" some threshold (see Chapter 13). The two most common examples in hydrology are constituent concentrations less than the detection limit and floods which are known to be less than some threshold of perception (e.g. the annual flood of 1887 was not sufficiently large that local record keepers bothered to record the maximum stage). The existence of censored values complicates the use of the previously discussed parametric procedures and all of the procedures involving removal of the effect of an exogenous variable. Any arbitrary choice of a value to represent the censored values (e.g., zero or the reporting limit) can give inaccurate results for hypothesis tests and biased estimates of trend slopes (Helsel, 1990).

A parametric approach to the detection of trends in censored data is the estimation of the parameters of a linear regression model relating Y to T, or Y to T and X, through the method of maximum likelihood estimation (MLE), also referred to as Tobit estimation (Hald, 1949; Cohen, 1950). These effects can be modeled simultaneously in this approach as can be done in a conventional multiple regression. Because the MLE method assumes a linear model with normally distributed errors, transformations (such as logarithms) of Y and X are frequently required to make the data more nearly normal and improve the fit of the MLE regression. Failure of the data to conform to these assumptions will tend to lower the statistical power of the test, and give unreliable estimates of the model parameters. The Type I error of the test is, however, relatively insensitive to violations of the normality assumption.

An extension of the MLE method was developed by Cohen (1976) to provide estimates of regression model parameters for data records with multiple censoring levels. An adjusted MLE method for multiply-censored data that is less biased in certain applications than the MLE method of Cohen (1976) was also recently developed by Cohn (1988). The availability of multiply-censored MLE methods is noteworthy for the analysis of lengthy water-quality records with censored values since these records frequently have multiple reporting limits that reflect improvements in the accuracy of analytical methods (and reductions in reporting limits) with time. Similarly, the multiply-censored case can arise in flood studies in that some very old portion of a flood record may contain estimates of only the very largest floods while a more recent part of the record (when flood plain development was more intense and record keeping more complete) may contain estimates of floods exceeding a more moderate threshold.

The Mann-Kendall test can be used without any difficulty when only one censoring threshold exists. Comparisons between all pairs of observations are possible. All the "less thans" are less than the other values and are considered to be tied with each other. Thus the S statistic and τ are easily computed using the tie correction for the standard deviation (see Chapter 8) in the large-sample approximation. Equation 8.4 for the corrected standard deviation is repeated here.

$$\sigma_S = \sqrt{\frac{[n\,(n-1)\,(2n+5) - \sum_{i=1}^{n} t_i\,(i)\,(i-1)\,(2i+5)\,]}{18}} \qquad [12.10]$$

When more than one detection limit exists, the Mann-Kendall test can not be performed without further censoring the data. Consider the data set: <1, <1, 3, <5, 7. How can a <1 and <5 be compared? A 3 and a <5? These ambiguities make the test impossible to compute. The only way to perform a Mann-Kendall test is to censor and recode the data at the highest detection limit. Thus, these data would become: <5, <5, <5, <5, 7. There is certainly a loss of information in making this change, and a loss of power to detect any trends which may exist. Other nonparametric tests incorporating multiple detection limits are briefly discussed in chapter 13. None can be performed when the change in detection limits is a function of the process being tested for, such as time. When this occurs, the change in censoring limit over time will induce the test to find a change in Y over time more often than it actually occurs (α will actually be larger than that which was stated for the test).

Although the <u>sign</u> of the estimated Theil trend slope is accurate for highly censored data records, the <u>magnitude</u> of the slope estimate is likely to be in error. Substitution of an arbitrarily chosen value between zero and the reporting limit for all censored values will likely give biased estimates of the trend slope. While the amount of bias cannot be stated precisely, the presence of only a few nondetected values in a record (less than about five percent) is not likely to affect the accuracy of the trend slope magnitude significantly.

Table 12.8 classifies monotonic trend tests for data sets with censored data.

	Not Adjusted for X	**Adjusted for X**
Nonparametric	Mann-Kendall test for trend on Y	No test yet available
Fully Parametric	Tobit regression of Y on T	Tobit regression of Y on X and T

Table 12.8 Classification of monotonic trend tests for censored data

Exercises

12.1 During the period 1962-1969 the Green River Dam was constructed about 35 miles upstream of a gaging station on the Green River at Munfordville, Kentucky. It is a flood control dam, thus it regulates flow (changes the flow duration curve) but has little or no effect on total annual runoff from this 1660 square-mile basin.

The question is this--over the period of record 1952-1972 (which includes pre-dam, transition, and regulated periods) has there been a monotonic trend in sediment transport? The data available in Appendix C17 are the year, suspended sediment load in thousands of tons per year, and the annual discharge in cfs-days.

Using each of the four trend analysis approaches described in the chapter, what would you conclude about suspended sediment trends?

12.2: Seasonal Kendall Test with censored data

The following data are dissolved lead concentrations (in mg/L) for the Potomac River at Chain Bridge at Washington, D.C.

	Winter	Spring	Summer	Fall
1973	–	4	3	–
1974	–	3	2	6
1975	4	5	5	11
1976	2	3	6	4
1977	16	2	18	17
1978	26	7	4	–
1979	5	3	9	6
1980	<2	<2	<2	<2
1981	5	<2	<2	<2
1982	<2	2	–	2
1983	2	3	<2	5
1984	3	5	<2	<2
1985	<2	3	–	<2

Compute Kendall's S and Var [S] for each season and perform a Mann-Kendall test for each season, reporting p-values. Then compute the Seasonal Kendall statistic and its variance and report the p-value for the Seasonal Kendall test.

12.3 The Maumee River, Ohio is a major tributary to Lake Erie. The control of phosphorus inputs to the lake has been a major concern since the early 1970's. In Appendix C18 are 133 observations of instantaneous load and streamflow for the Maumee River from the 1972 through 1986. The variables in the date set are TIME (decimal time in years), MONTH (month of the year), Q (discharge in thousands of cfs), and LOAD (total phosphorus load in tons per day).

Using the various approaches listed in this chapter, describe the trend in the data set. Try each of the approaches listed in table 12.2. For the nonparametric approaches compute the seasonal-Kendall test.

Then using the model you select as most appropriate, estimate loads for the following four cases:

(Q , TIME) = (11 , 1972.5), (1 , 1972.5), (11 , 1986.5), (1 , 1986.5).

12.4 Major ion chemistry of groundwater may be altered as it remains in contact with aquifer materials. One common alteration is ion exchange of calcium with sodium. Carbon-14 can be used as a measure of the age of groundwater, with C-14 (modern carbon) decreasing with increasing age. Compute the p-value and determine whether there is a trend to lower calcium concentrations with increasing age for the following data:

%Carbon-14	Calcium, in % meq	%Carbon-14	Calcium, in % meq
(old) 12.4	20.16	48.9	16.56
19.5	0.46	52.0	22.88
23.9	4.18	55.6	12.64
25.3	28.95	57.7	14.26
28.2	20.00	58.1	13.37
31.5	0.57	66.2	35.22
33.1	1.84	67.1	24.43
33.4	13.99	(young) 71.8	60.24
38.6	18.02		

12.5 A well screened in the middle aquifer in New Jersey is investigated to determine if groundwater levels have declined over time as a result of pumpage. The data are given in Appendix C19 (A. Pucci, written communication, U.S. Geological Survey, Trenton, NJ). Conduct a trend analysis to determine whether a decline has occurred. Estimate the rate of decline over the period of record.

Chapter 13

Methods for Data Below the Reporting Limit

To comply with environmental regulations, an industry must show that the daily mean copper concentration in its wastewater discharge does not exceed the legal standard. Yet for many of the wastewater samples taken at two hour intervals, concentrations are below the analytical reporting limit of the laboratory. These "less-thans" make it impossible to compute a simple mean concentration. When the industry substitutes a zero for each less-than, the standard is not exceeded. When the regulatory agency substitutes a value equal to the reporting limit, the standard is exceeded. Which is correct? Has the law been violated?

Ground-water quality is measured both upgradient and downgradient of a waste-disposal site. Comparison of the two groups of data is performed to determine whether contamination of the ground-water system has occurred. Usually t-tests are employed for this purpose, and yet the t-test requires estimates of means and standard deviations which are impossible to obtain unless numerical values are fabricated to replace any less-thans present in the data. By substituting one number, the two groups appear the same. Substituting a second number causes H_0 to be rejected, and the two groups to be declared different. Which is correct?

As trace substances in the world's soils, air and waters are increasingly investigated, concentrations are more frequently being encountered which are less than limits deemed reliable enough to report as numerical values. These less-than values -- values stated only as "<rl", where rl is called the "reporting limit" or "detection limit" or "limit of quantitation" (Keith et al., 1983) -- present a serious interpretation problem for data analysts. Estimates of summary statistics which best represent the entire distribution of data, both below and above the reporting limit, are necessary to accurately analyze environmental conditions. Also needed are hypothesis test and regression procedures that provide valid conclusions and models for such data. These needs must be met using the only information available to the data analyst: concentrations measured above one or more reporting limits, and the observed frequency of data below those limits.

This chapter discusses the most appropriate statistical procedures given that data have been reported as less-thans. It does not consider the alternative of reporting numerical values for all data, including those below reporting limits -- see ASTM (1983), Porter et al. (1988) and Gilliom et al. (1984) for discussion of this alternative.

13.1 Methods for Estimating Summary Statistics

Methods for estimating summary statistics of data which include less-thans (statisticians call these "censored data") can be divided into the three classes discussed below: simple substitution, distributional, and robust methods. Recent papers have documented the relative performance of these methods. Gilliom and Helsel (1986) and Gleit (1985) compared the abilities of several estimation methods in detail over thousands of simulated data sets. Helsel and Gilliom (1986) then applied these methods to numerous water-quality data sets, including those which are not similar to the assumed distributions required by the distributional methods. A single case study was reported by Newman and Dixon (1990). Helsel and Cohn (1988) dealt with censoring at multiple reporting limits. Large differences were found in these methods' abilities to estimate summary statistics for censored data.

Methods may be compared based on their ability to replicate true population statistics. Departures from true values are measured by the root mean squared error (RMSE), which combines both bias and lack of precision. The RMSE of the estimate of the mean $\bar{x}$ in comparison to the true population value μ is shown in equation 13.1. Similar equations would be used for estimation of other summary statistics.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(\bar{x}-\mu)^2}{\mu} \Big/ N} \qquad [13.1]$$

Methods whose estimates $\bar{x}$ are closer to the true value μ have lower RMSEs, and are considered better.

13.1.1 Simple Substitution Methods

Simple substitution methods (Figure 13.1) substitute a single value such as one-half the reporting limit for each less-than value. Summary statistics are calculated using both these fabricated numbers along with the values above the reporting limit. These methods are widely used, but have no theoretical basis. As Figure 13.1 shows, the distributions resulting from simple substitution methods have large gaps, and do not appear realistic.

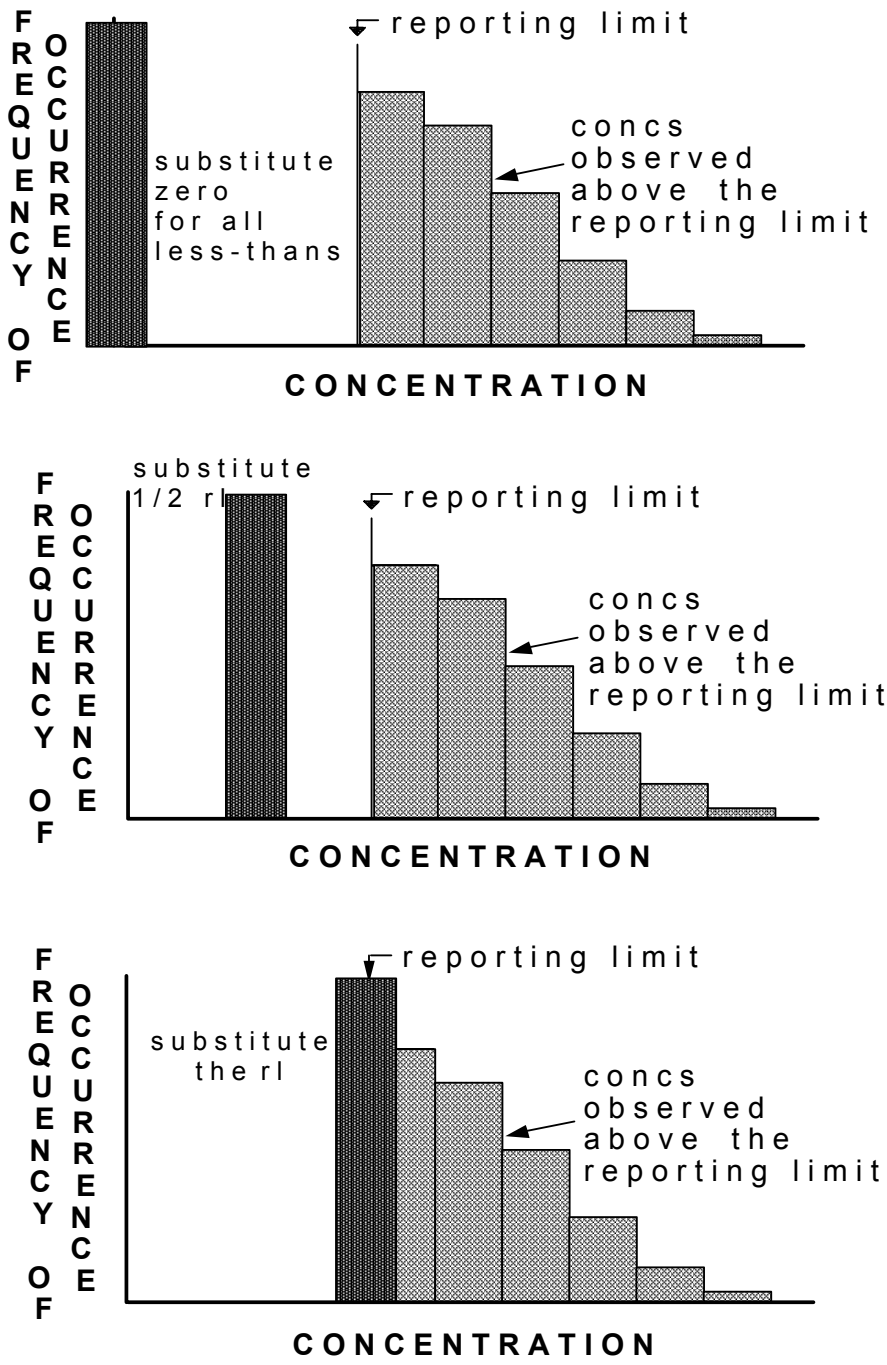


Figure 13.1. Histograms for simple substitution methods.

Studies cited above determined that simple substitution methods performed poorly in comparison to other procedures. Substitution of zero produced estimates of mean and median which were biased low, while substituting the reporting limit resulted in estimates above the true value. Results for the standard deviation and IQR, and for substituting one-half the reporting limit, were also far less desirable than alternative methods. With the advent of powerful desktop

computers to perform more complex calculations there appears to be no reason to use simple substitution methods. As the choice of value to be substituted is essentially arbitrary without some knowledge of instrument readings below the reporting limit, and as large differences may occur in the resulting estimates, simple substitution methods are not defensible.

13.1.2 Distributional Methods

Distributional methods (Figure 13.2) use the characteristics of an assumed distribution to estimate summary statistics. Data both below and above the reporting limit are assumed to follow a distribution such as the lognormal. Given a distribution, estimates of summary statistics are computed which best match the observed concentrations above the reporting limit and the percentage of data below the limit. Estimation methods include maximum-likelihood estimation or MLE (Cohen, 1959), and probability plotting procedures (Travis and Land, 1990). MLE estimates are more precise (lower RMSE) than probability plotting, and both methods are unbiased, when observations fit the assumed distribution exactly and when the sample size is large. However, this is rarely the case in environmental studies. When data do not match the observed distribution, both methods may produce biased and imprecise estimates. Thus the most crucial consideration when using distributional methods is how well the data can be expected to fit the assumed distribution. Even when distributional assumptions are correct, MLEs have been shown to produce estimates with large bias and poor precision for the small sample sizes of n=5, 10, and 15 (Gleit, 1985). MLE methods are commonly used in environmental disciplines such as air quality (Owen and DeRouen, 1980) and geochemistry (Miesch, 1967).

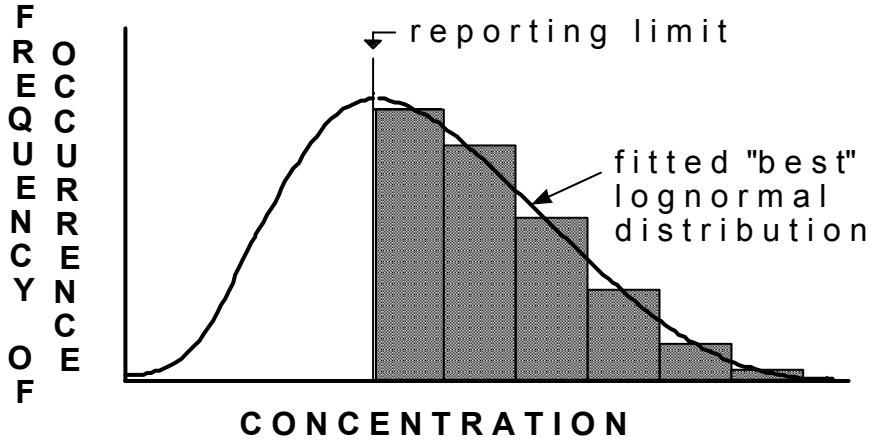
Assuming a lognormal distribution for concentrations, MLEs for larger (n=25, 50) data sets have provided excellent estimates of percentiles (median and IQR) for a variety of data distributions realistic for environmental studies, even those which are not lognormal. However, they have not worked as well for estimating the mean and standard deviation (Gilliom and Helsel, 1986). There are two reasons why this is so.

First, the lognormal distribution is flexible in shape, providing reasonable approximations to data which are nearly symmetric and to some positively-skewed distributions which are not lognormal. Thus the lognormal can mimic the actual shape of the data over much of the distribution, adequately reproducing percentile statistics even though the data were not truly lognormal in shape. However, the moment statistics (mean and standard deviation) are very sensitive to values of the largest observations. Failure of the assumed distribution to fit these observations will result in poor estimates of moments.

Second, there is a transformation bias in lognormal MLE inherent in computing estimates of the mean and standard deviation for any transformation -- including logarithms -- and then retransforming back to original units. Compensating for this bias often requires an assumption about distributional shape. In Chapter 9 transformation bias was discussed in the context of

regression. The same phenomena is present for estimates of the mean. Indeed, if no explanatory variables are significant then a regression model simplifies to estimating the mean. Percentiles, however, can be directly transformed between measurement scales without bias.

Maximum Likelihood (MLE) -- fits 'best' lognormal distribution to the data, and then



determines summary statistics of the fitted distribution to represent the data.

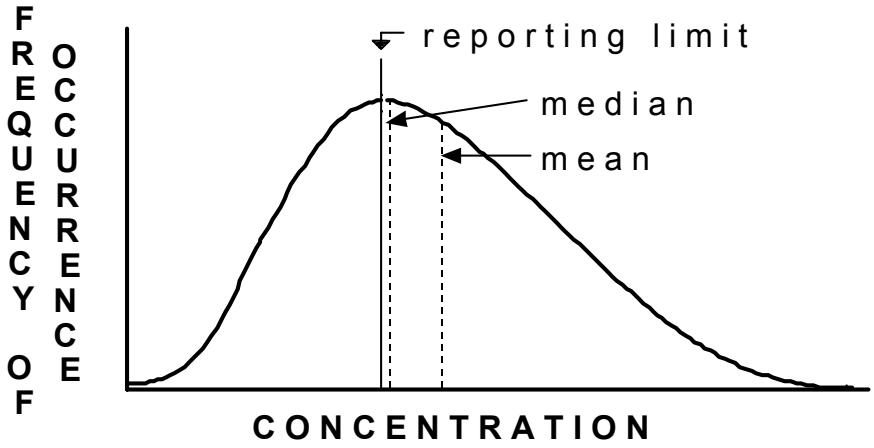


Figure 13.2. Distributional (MLE) method for computing summary statistics.

Two less-frequently used distributional methods are a "fill-in with expected values" MLE technique (Gleit, 1985) and a probability plot method which estimates the mean and standard deviation by the intercept and slope, respectively, of a line fit to data above the reporting limit (Travis and Land, 1990). Probability plot methods are easy to compute with standard statistics software, an advantage for practitioners. Both methods suffer from transformation bias when estimates are computed in one scale and then retransformed back into original units. Thus

Travis and Land (1990) recommended the probability plot for estimating the geometric mean. Its use for estimating the mean in original units would have to take transformation bias into consideration. Both methods should be somewhat less precise than MLEs.

13.1.3 Robust Methods

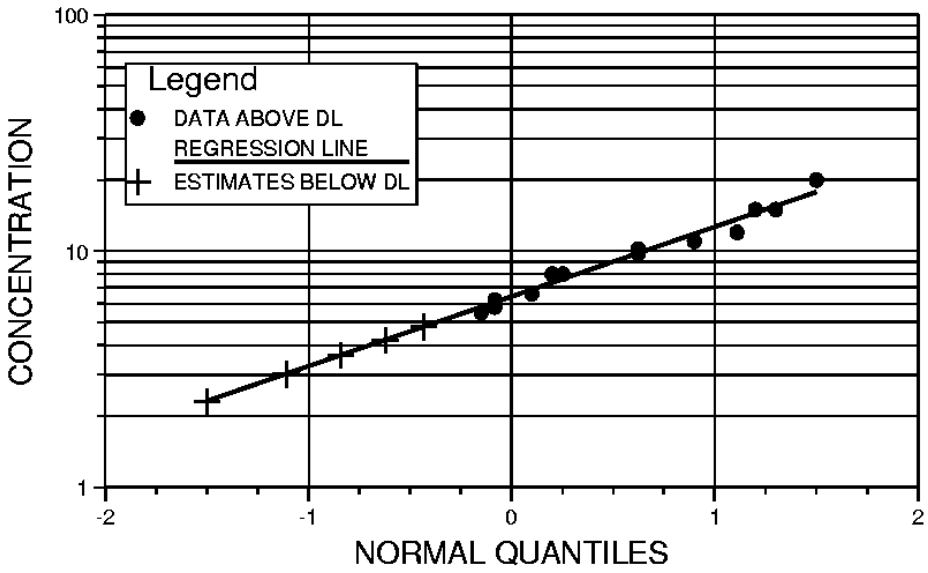
Robust methods (Figure 13.3) combine observed data above the reporting limit with below-limit values extrapolated assuming a distributional shape, in order to compute estimates of summary statistics. A distribution is fit to the data above the reporting limit by either MLE or probability plot procedures, but the fitted distribution is used only to extrapolate a collection of values below the reporting limit. These extrapolated values are <u>not</u> considered as estimates for specific samples, but only used collectively to estimate summary statistics. The robustness of these methods result primarily from their use of observed data rather than a fitted distribution above the reporting limit. They also avoid transformation bias by performing all computations of summary statistics in original units.

Robust methods have produced consistently small errors for all four summary statistics in simulation studies (Gilliom and Helsel, 1986), as well as when applied to actual data (Helsel and Gilliom, 1986). Robust methods have at least two advantages over distributional methods for computation of means and standard deviations. First, they are not as sensitive to the fit of a distribution for the largest observations because actual observed data are used rather than a fitted distribution above the reporting limit. Second, estimates of extrapolated values can be directly retransformed and summary statistics computed in the original units, avoiding transformation bias.

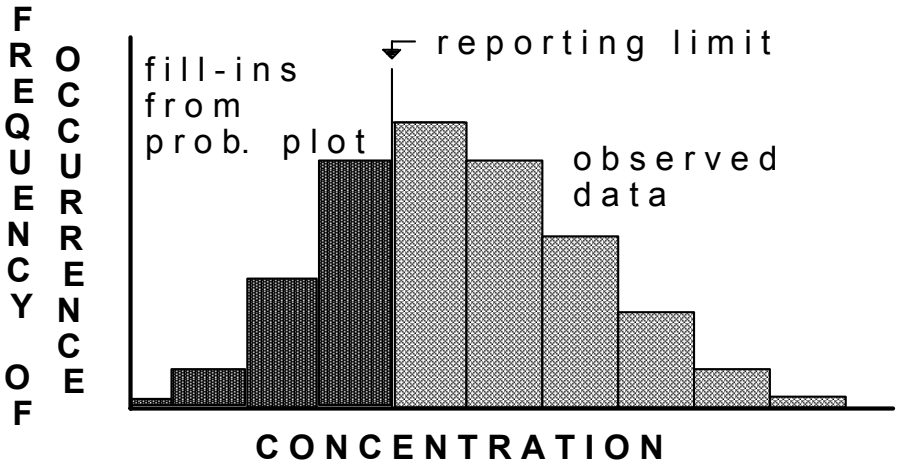
13.1.4 Recommendations

In practice, the distributions of environmental data are rarely if ever known, and may vary between constituents, time periods, and locations. Robust methods should therefore be used to protect against the possibly large errors of distributional methods when estimating the mean and standard deviation. Either robust probability plot or distributional MLE procedures have been shown to perform well for estimating the median and IQR. Use of these methods rather than simple substitution methods for environmental data should substantially lower estimation errors for summary statistics.

As an alternative to estimating percentiles, sample values can sometimes be used. When less than 50% of the data are below the reporting limit, the sample median is known. Similarly, when less than 25% of the data are censored, the sample IQR is known. Some information is available about percentiles when even larger amounts of data lie below the reporting threshold, as shown in the examples below. Unfortunately no similar process is available for sample estimates of mean and standard deviation.



A.



B.

Figure 13.3. Robust (probability plot) method of estimating summary statistics

A) regression of log of concentration versus normal score is used to extrapolate "fill-in" values below the reporting limit.

B) these "fill-ins" are retransformed back to original units, and combined with data above the reporting limit to compute estimates of summary statistics

Example 1

<1 <1 <1 <10 <10 <10 <50.

The mean and std deviation cannot be estimated by any method, as there are no data above the reporting limit. For the median and IQR, a great deal of information is present. To compute a median where all data are below one or more reporting limits and the sample size is odd, remove the < sign, compute the sample median, and then restore the < sign. Therefore the median is <10. The IQR must equal the sample 75th percentile, as the 25th percentile could equal zero. Here the IQR is <10.

Example 2:

<1 <1 <1 <10 <20 <20

When all data are below one or more reporting limits and n is even, again remove the < signs. The larger of the two center observations (the ([n/2]+1)th observation) is used as the median, rather than the average of the two center observations as for uncensored data. Restore the < sign. The median of the above 6 points is <10. The IQR is computed as in example 1, and here would be <20.

Example 3:

<1 <1 <1 5 7 8 12 16 25

For data above and below one reporting limit, the sample median is known to be 7, as less than 50% of the data are censored. Because more than 25% of the data are censored, the sample IQR must be computed as a range. If all the <1's are actually 0, the IQR = 14 – 0 = 14. If all <1's are very close to 1, the IQR = 13. So the sample IQR could be reported as "13 to 14" if that were of sufficient precision. Otherwise, the probability plot and maximum likelihood methods must be used to estimate the moment and percentile statistics.

13.1.5 Multiple Reporting Limits

Data sets may contain values censored at more than one reporting limit. This commonly occurs as limits are lowered over time at a single lab, or when data having different reporting limits are combined from multiple laboratories. Estimation methods belonging to the above three classes are available for this situation. A comparison of these methods (Helsel and Cohn, 1988) concluded that robust methods again provide the best estimates of mean and standard deviation, and MLEs for percentiles. For example, in Figure 13.4 the error rates for six estimation methods are compared to the error that would occur had all data been above the reporting limit (shown as the 100% line). Figure 13.5 shows the same information when the data differ markedly from a lognormal distribution. The simple substitution methods (ZE, HA and DL: substitution of zero, one-half and one times the reporting limit, respectively) have more error in most cases than does the robust probability plot method MR. Where the substitution methods

have lower RMSE, it is an artifact of constant, strongly biased estimates, also not a desirable result. The maximum likelihood procedure MM and the MLE adjusted for transformation bias AM show themselves to be excellent estimation methods for percentiles, but suffer from large errors when estimating the mean and standard deviation.

In summary, use of MLE for estimation of percentiles, and the robust probability plot method for estimating the mean and standard deviation, should greatly decrease errors as compared to simple substitution methods for data with multiple reporting limits.

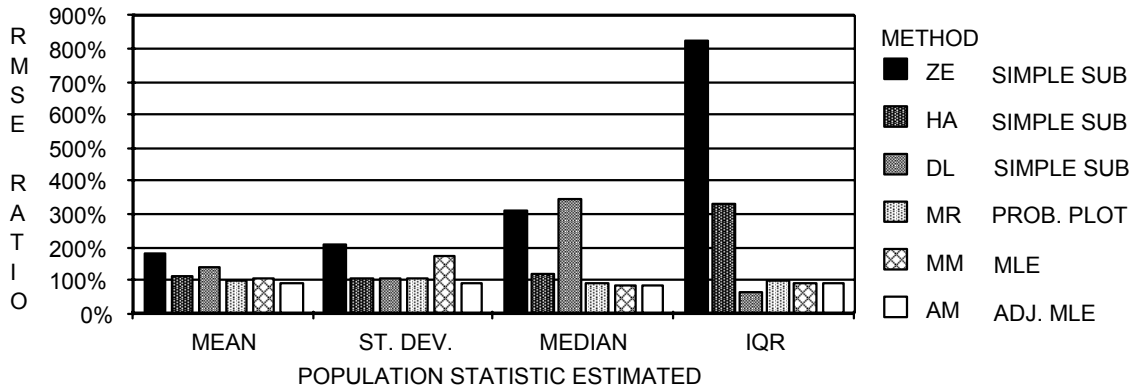


Figure 13.4. Error rates (RMSE -- root mean square error) of six multiple-detection methods divided by error rates for uncensored data estimates, in percent, for data similar to a lognormal distribution (from Helsel and Cohn, 1988)

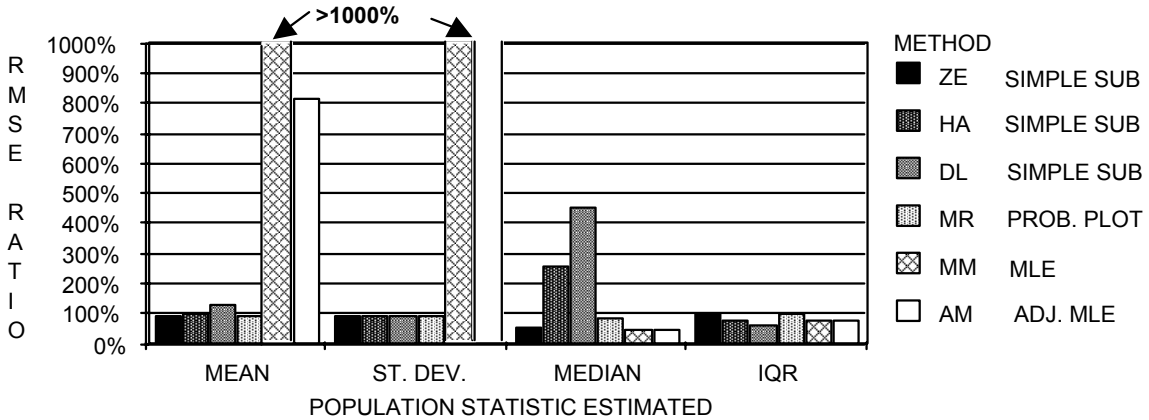


Figure 13.5. Error rates (RMSE -- root mean square error) of six multiple-detection methods divided by error rates for uncensored data estimates, in percent, for data not similar to a lognormal distribution. (from Helsel and Cohn, 1988)

Example 4:

For data above and below multiple reporting limits, such as

<1 <1 <1 5 7 8 <10 <10 <10 12 16 25

it is unclear whether the <10's are below or above 1, or 5, etc. Therefore ordering the data from smallest to largest is impossible. Instead, the probability plot method is used to compute the mean and standard deviation, and the maximum likelihood method for the median and IQR (Helsel and Cohn, 1988). These give the following:

mean	=	7.8	median	=	2.8
std dev	=	6.9	IQR	=	7.5.

13.2 Methods for Hypothesis Testing

Methods for hypothesis testing of censored data can also be classified into the three types of procedures: simple substitution, distributional or parametric, and robust or nonparametric. The advantages and disadvantages of each are summarized below.

13.2.1 Simple Substitution Methods

When censoring is present, values are often fabricated in order to perform parametric tests such as t-tests. Problems caused by such substitution methods are illustrated below. Investigators have also deleted censored data prior to hypothesis testing. This latter approach is the worst procedure, as it causes a large and variable bias in the parameter estimates for each group. After deletion, comparisons made are between the upper X% of one group versus the upper Y% of another, where X and Y may be very different. Such tests have little or no meaning.

Example 5

As an example of hypothesis test methods for censored data, tests will be performed to determine whether or not means or medians significantly differ between two groups. Two data sets were generated from lognormal distributions having the same variance, but with differing mean values. Sample statistics for the two data sets before and after censoring are given in Table 13.1. Prior to any censoring, group means were found to be different by a t-test ($p=0.04$, Table 13.2). The data were then censored at a reporting limit of 1 μg/L, so that all data below 1.0 were recorded as <1. This produced 14 less-than values (70%) in group A, and 5 less-than values (23%) in group B.

The simple substitution method for comparing two groups of censored data is to fabricate data for all less-than values, and include these "data" with detected observations when performing a t-test. No *a priori* arguments for fabrication of any particular value between 0 and the reporting limit can be made. Substituting zero for all less-than values, the means are declared significantly different ($p = 0.01$). Yet when the reporting limit of 1.0 is substituted, the means are not found to be different ($p = 0.19$). The conclusion is thus strongly dependent on the value substituted!

Fabrication of data followed by a t-test is an arbitrary process leading to ambiguous conclusions. It should be avoided.

13.2.2 Distributional Test Procedures

Parametric tests are also available which do not require substitutions for less-thans. Instead, maximum likelihood methods are used to solve the relevant equations. Where the distributional assumptions are appropriate, these relatively unknown tests have great utility.

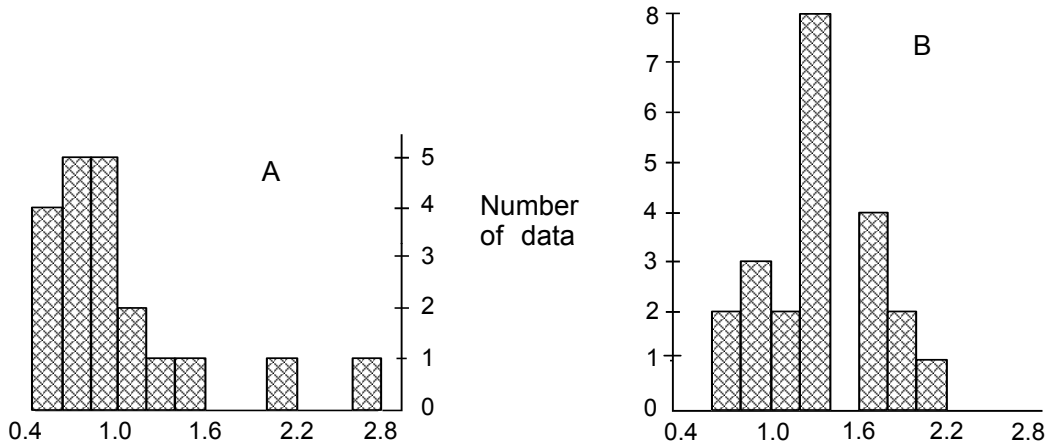
The distributional method for a t-test situation is performed using a regression procedure for censored data known as tobit regression (Judge et al, 1985). Tobit regression uses both the data values above the reporting limit, and the proportion of data below the reporting limit, to compute a slope coefficient by maximum likelihood. For a two-group test, the explanatory variable in the regression equation is the binary variable of group number, so that data in one group have a value of 0, and in the other group a value of 1. The regression slope then equals the difference between the two group means, and the t-test for whether this slope differs from zero is also a test of whether the group means differ. Tobit regression is discussed further in section 13.3. One advantage to Tobit regression for hypothesis testing is that multiple reporting limits may easily be incorporated. The caution for its use is that proper application does require the data in both groups to be normally distributed around their group mean, and for the variance in each group to be equal. For large amounts of censoring these restrictions are difficult to verify.

13.2.3 Nonparametric Tests

With nonparametric tests, no fabrication of data values is required. All censored data are represented by ranks which are tied at values lower than the lowest number above the reporting limit. The rank-sum test compares the medians of two independent data groups (Chapter 5).

Prior to censoring, a rank-sum test on the example 5 data produced a much lower p-value ($p=0.003$) than did the t-test (Table 13.2). This lower p-value is consistent with the proven greater power of the nonparametric test to detect differences between groups of skewed data, as compared to the t-test. To compute the rank-sum test on censored data, the 19 less-than values are considered tied at the lowest value, with each assigned a rank of 10 (the mean of ranks 1-19). The next highest value, the data point just above the reporting limit, obtains a rank of 20. All data above the reporting limit will have ranks identical to those which would have been obtained had no censoring been present. The resulting p-value is 0.002, essentially the same as for the uncensored data, and the two groups are easily declared different. Thus in this example the nonparametric method makes very efficient use of the information contained in the less-than values, avoids arbitrary assignment of fabricated values, and accurately represents the lack of knowledge below the reporting limit. Results do not depend on any distributional assumption.

TABLE 13.1--Characteristics of Two Lognormal Data Groups (A and B)



A		B
20	no. observations	22
1.00	mean	1.32
0.57	std. deviation	0.39
0.65	25th percentile	1.07
0.85	median	1.25
1.07	75th percentile	1.66
14	no. <rl	5

TABLE 13.2-- Significance Tests Between Groups A and B

Hypothesis test used	test statistic	p
Uncensored data		
t-test (Satterthwaite approx.)	–2.13	0.04
regression with binary variable	–2.17	0.04
rank-sum test		
	–2.92	0.003
After imposing artificial reporting limit		
t-test		
less-thans = 0.0	–2.68	0.01
less-thans = 0.5	–2.28	0.03
less-thans = 1.0	–1.34	0.19
tobit regression with binary variable	–2.28	0.03
rank-sum test	–3.07	0.002

When severe censoring (near 50% or more) occurs, all of the above tests will have little power to detect differences in central values. The investigator will be inhibited in stating conclusions about the relative magnitudes of central values, and other characteristics must be compared. For example, contingency tables (Chapter 14) can test for a difference in the proportion of data

above the reporting limit in each group. The test can be used when the data are reported only as detected or not detected. It may also be used when response data can be categorized into three or more groups, such as: below detection, detected but below some health standard, and exceeding standards. The test determines whether the proportion of data falling into each response category differs as a function of different explanatory groups, such as different sites, land use categories, etc.

13.2.4 Hypothesis Testing With Multiple Reporting Limits

More than one reporting limit is often present in environmental data. When this occurs, hypothesis tests such as comparisons between data groups are greatly complicated. It can be safely said that fabrication of data followed by computation of t-tests or similar parametric procedures is at least as arbitrary with multiple reporting limits as with one reporting limit, and should be avoided. The deletion of data below all reporting limits prior to testing should also be <u>completely</u> avoided.

Tobit regression can be utilized with multiple reporting limits. Data should have a normal distribution around all group means and equal group variances to use the test. These assumptions are difficult to verify with censored data, especially for small data sets.

One robust method which can always be performed is to censor all data at the highest reporting limit, and then perform the appropriate nonparametric test. Thus the data set

<1 <1 <1 5 7 8 <10 <10 <10 12 16 25

would become <10 <10 <10 <10 <10 <10 <10 <10 <10 12 16 25.

and a rank-sum test performed to compare this with another data set. Clearly this produces a loss of information which may be severe enough to obscure actual differences between groups (a loss of power). However, for some situations this is the best that can be done.

Alternatively, nonparametric score tests common in the medical "survival analysis" literature can sometimes be applied to the case of multiple reporting limits (Millard and Deverel, 1988). These tests modify uncensored rank test statistics to compare groups of data. The modifications allow for the presence of multiple reporting limits. In the most comprehensive review of these score tests, Latta (1981) found most of them to be inappropriate for the case of unequal sample sizes. The Peto-Prentice test with asymptotic variance estimate was found by Latta (1981) to be the least sensitive to unequal sample sizes. Another crucial assumption of score tests is that the censoring mechanism must be independent of the effect under investigation (see box). Unfortunately, this is often not the case with environmental data.

Examples when a score test would be inappropriate.

Score tests are inappropriate when the censoring mechanism differs for the two groups. That is, the probability of obtaining a value below a given reporting limit differs for the two groups when the null hypothesis that the groups are identical is true.

1. Suppose a trend over time was being investigated. The first five years of data were produced with a method having a reporting limit of 10 mg/L; the second five years used an improved method with 1 mg/L as its reporting limit. A score test of the first half of the data versus the second would not be valid, as the censoring mechanism itself varied as a direct function of time.
2. Two groups of data are compared as in a rank-sum test, but most of the data from group A were measured with a chemical method having 1 as its reporting limit, while most of group B were measured with a method having 10 as its reporting limit. A score test would not yield valid results, as the censoring mechanism varies as a function of what is being investigated (the two groups).

Examples when a score test would be appropriate.

A score test yields valid results when the change in censoring mechanism is not related to the effect being measured. Stated another way, the probability of obtaining data below each reporting limit is the same for all groups, assuming the null hypothesis of no trend or no difference is true. Here a score test provides much greater power than artificially censoring all data below the highest reporting limit before using the rank-sum test.

1. Comparisons were made between two groups of data collected at roughly the same times, and analyzed by the same methods, even though those methods and reporting limits changed over time. Score tests are valid here.

2. Differing reporting limits resulted from analyses at different laboratories, but the labs were assigned at random to each sample. Censoring is thus not a function of what is being tested, but is a random effect, and score tests would be valid.

13.2.5 Recommendations

Robust hypothesis tests have several advantages over their distributional counterparts when applied to censored data. These advantages include: (1) the ability to disregard whether data adhere to a normal distribution. Verifying normality is difficult to do with censored data; (2) greater power for the skewed distributions common to environmental data; and (3) data below the reporting limit are incorporated without fabrication of values or bias. Information contained in less-than values is accurately used, not misrepresenting the state of that information.

Tests incorporating multiple reporting limits are more problematic, and should be an area of future research. When adherence to a normal distribution can be documented, Tobit regression

offers the ability to incorporate multiple reporting limits in a distributional test regardless of a change in censoring mechanism. Nonparametric score tests require consistency in censoring mechanism with respect to the effect being tested.

13.3 Methods For Regression With Censored Data

With censored data the use of ordinary least squares (OLS) for regression is prohibited. Coefficients for slopes and intercept cannot be computed without values for the censored observations, and substituting fabricated values may produce coefficients strongly dependent on the values substituted. Four alternative methods capable of incorporating censored observations are described below.

The choice of method depends on the amount of censoring present, as well as on the purpose of the analysis. For small amounts of censoring (below 20%), either Kendall's line or the tobit line may be used. Kendall's would be preferred if the residuals were not normally distributed, or when outliers are present. For moderate censoring (20-50%), Tobit or logistic regression must be used. With large amounts of censoring, inferences about concentrations themselves must be abandoned, and logistic regression employed. When both the explanatory and response variables are censored, tobit regression is applicable for small amounts of censoring. For larger amounts of censoring, contingency tables or rank correlation coefficients are the only option.

13.3.1 Kendall's Robust Line Fit

When one censoring level is present, Kendall's rank-based procedure for fitting a straight line to data can test the significance of the relationship between a response and explanatory variable (Chapter 10). An equation for the line, including an estimate of the slope, is usually also desirable. This can be computed when the amount of censoring is small. Kendall's estimate of slope is the median of all possible pairwise slopes of the data. To compute the slope with censoring, twice compute the median of all possible slopes, once with zero substituted for all less-thans, and once with the reporting limit substituted. For small amounts of censoring the resulting slope will change very little, or not at all, and can be reported as a range if necessary. If the slope value change is of an unacceptable magnitude, tobit or logistic regression must be performed.

Research is underway on methods based on scores similar to those for hypothesis tests with multiply-censored data that may allow robust regression fits to data with multiple reporting limits (McKean and Sievers, 1989).

13.3.2 Tobit Regression

Censored response data can be incorporated along with uncensored observations into a procedure called tobit regression (Judge et al., 1985). It is similar to OLS except that the coefficients are fit by maximum-likelihood estimation. MLE estimates of slope and intercept are

based on the assumption that the residuals are normally distributed around the tobit line, with constant variance across the range of predicted values. Again, it is difficult to check these assumptions with censored data. Should the data include outliers, these can have a strong influence on the location of the line and on significance tests (Figure 13.6), as is true with uncensored OLS. Verification of linearity and constant variance assumptions should be attempted when only small amounts of data are censored using residuals plots. Residuals for uncensored observations would be plotted versus predicted values. For larger percentages of less-thans, decisions whether to transform the response variable must often be made based on previous knowledge ("metals always need to be log-transformed", etc.). Tobit regression is also applicable when both the response and explanatory variables are censored, such as a regression relationship between two chemical constituents. However, the amount of censoring must be sufficiently small that the linearity, constant variance, and normality assumptions of the procedure can be checked. Finally, Cohn (1988) as well as others have proven that the tobit estimates are slightly biased, and have derived bias corrections for the method.

13.3.3 Logistic Regression

Here the response variable is categorical. No longer is a concentration being predicted, but a probability of being in discrete binary categories such as above or below the reporting limit. One response (above, for example) is assigned a value of 1, and the second response a 0. The probability of being in one category versus the second is tested to see if it differs as a function of continuous explanatory variable(s). Examples include predicting the probability of detecting

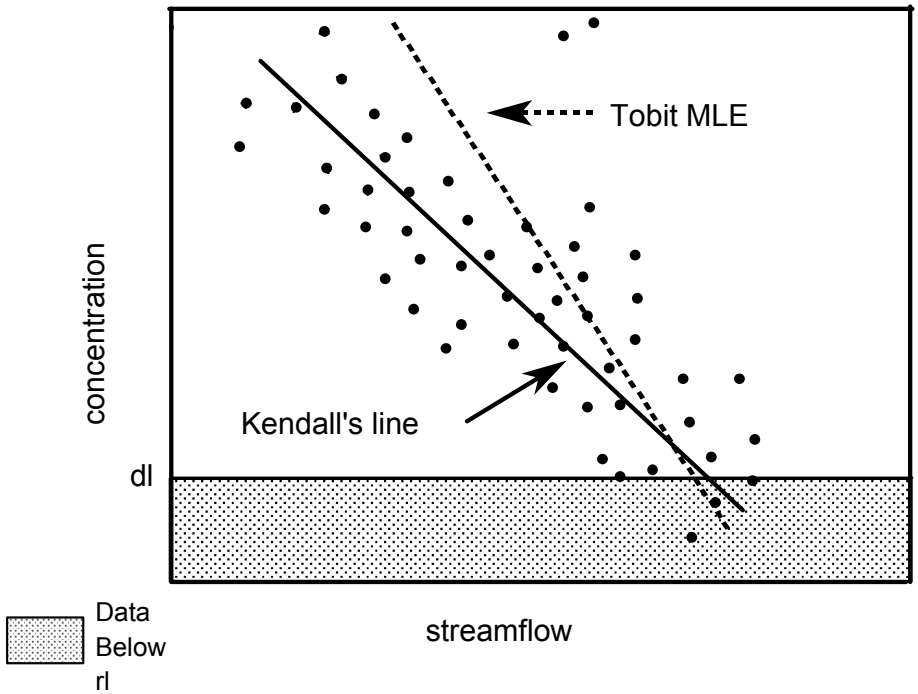


Figure 13.6. Kendall's and tobit MLE lines for censored data with outliers. Note the tobit line is strongly influenced by outliers.

concentrations of some organic contaminant from continuous variables such as nitrate concentrations, population density, percent of some appropriate land use variable, or of irrigation intensity. Predictions from this regression-type relationship will fall between 0 and 1, and are interpreted as the probability [p] of observing a response of 1. Therefore [1−p] is the probability of a 0 response.

Logistic regression may be used to predict the probabilities of more than 2 response categories. When there are m>2 ordinal (may be placed in an order) responses possible, (m−1) equations must be derived from the data. For example, if 3 responses are possible (concentrations below rl =0, above rl but below health standards =1, and above health standards =2), two logistic regressions must be computed. First, an equation must be written for the probability of being nonzero (the probability of being above the rl). Next the probability of a 2 (probability of being above the health standard) is also modelled. Together, these two equations completely define the three probabilities p(y=0), p(y=1), and p(y=2) as a function of the explanatory variables. See Chapter 15 for more detail.

13.3.4 Contingency Tables

Contingency tables are useful in the regression context if both explanatory and response variables contain censoring. For example, suppose the relationship between two trace metals in soils (such as arsenic and aluminum) is to be described. The worst procedure would again be to throw away the data below the reporting limits, and perform a regression. Figure 13.7 shows that a true linear relationship with negative slope could be completely obscured if censored data were ignored, and only data in the upper right quadrant investigated. Contingency tables provide a measure of the strength of the relationship between censored variables -- the phi statistic ϕ (Chapter 14), a type of correlation coefficient. An equation which describes this relationship, as per regression, is not available. Instead, the probability of y being in one category can be stated as a function of the category of x. For the Figure 13.7 data, the probability of arsenic being above the reporting limit is 21/36 = 0.58 when aluminum is above reporting limit, and 17/18 = 0.94 when aluminum is below the reporting limit.

13.3.5 Rank Correlation Coefficients

The robust correlation coefficients Kendall's tau or Spearman's rho (Chapter 8) could also be computed when both variables are censored. All values below the reporting limit for a single variable are assigned tied ranks. Rank correlations do not provide estimates of the probability of exceeding the reporting limit as does a contingency table. So they are not applicable in a regression context, but would be more applicable than contingency tables in a correlation context. One such context would be in "chemometrics" (Breen and Robinson, 1985): the computation of correlation coefficients for censored data as inputs to a principal components or factor analysis.

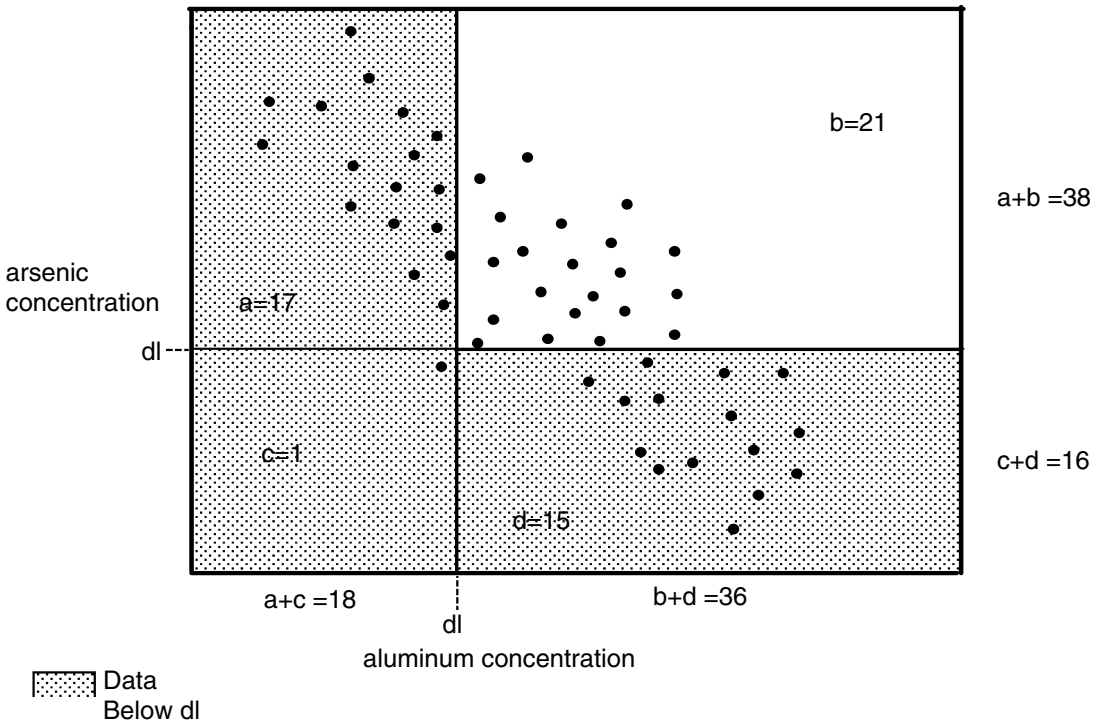


Figure 13.7. Contingency table relationship between two censored variables. (Ignoring censored data would produce the misleading conclusion that no relationship exists between the two variables)

13.3.6 Recommendations

Relationships between variables having data below reporting limits can be investigated in a manner similar to regression. Values should not be fabricated for less-thans prior to regression. Instead, Table 13.3 summarizes alternative methods appropriate for censored data. For small amounts of censoring and one reporting limit, Kendall's robust line can be fit to the data. For moderate censoring and/or multiple reporting limits, tobit regression can be performed. For more severe censoring of the dependent variable, logistic regression is appropriate. When both response and explanatory variables contain severe censoring, contingency tables and rank correlation coefficients can be performed.

Estimation of Summary Statistics		
Mean and Standard Deviation		Percentiles
Robust Probability Plot		Robust Probability Plot or MLE
	Hypothesis Tests	
	One Reporting Limit	Several Reporting Limits
Compare 2 groups:	rank-sum test	tobit regression
Compare >2 groups:	Kruskal-Wallis test	tobit regression
Severe Censoring (>50%):	above tests, or contingency tables	–
	Regression	
Small % censoring	Moderate % censoring	Large % censoring
Kendall's robust line	tobit regression	logistic regression
tobit regression	logistic regression	contingency tables

Table 13.3 Recommended Techniques for Interpretation of Censored Data

Exercises

13.1 Below are concentrations of triphenyltin (TPT) measured in a sediment core by Fent and Hunn (1991). Estimate the mean, standard deviaion, median and interquartile range for these data.

Concentrations of TPT, in g/kg dry weight

51	29	71	69	34	56	83	<2	<2	107	35	<2
26	4	10	<2	2	<2						

13.2 Below are depths (bottom of segment) for the 18 TPT concentrations of exercise 13.1. Compute the significance of the relationship between concentration and depth for this core.

Depth (cm) of bottom of sediment core

0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6
7	8	9	10	11	12						

13.3 Silver concentrations in standard solutions were reported by several laboratories in an inter-lab comparison [Janzer, 1986]. The 56 analyses included 36 values below one of 12 detection limits. One large outlier (a "far outside" value on a boxplot) of 560 µg/L was also reported. Compute the mean, standard deviation, median and interquartile range for the data presented below:

0.8	<25	<5	<0.2	<0.5	5.0	<0.3	<0.2	0.1
2.7	<0.1	<20	1.42.0	<2.5	2.0	2.0	<1	
<10	<1	<0.2	1.0<10	<0.2	0.2	1.2	<1	
1.0	<6	<1	0.7<1	<53.2	2.0	10.0		
1.0	4.4	<1	<1<1	<20	<5	<10	<10	
90	1.5	<1	<2<10	560<5	0.1	<20		
<1	<0.1							

Chapter 14

Discrete Relationships

Three aquifers are sampled to determine whether they differ in their concentrations of copper. In all three, over 40 percent of the samples were below the detection limit. What methods will test whether the distribution of copper is identical in the three aquifers while effectively incorporating data below the detection limit?

Counts of three macroinvertebrate species were measured in three stream locations to determine ecosystem health. The three species cover the range of tolerance to pollution, so that a shift from dominance of one species to another is an indication of likely contamination. Do the three locations differ in their proportions of the three species, or are they identical?

This chapter presents methods to evaluate the relationship between two discrete (also called categorical) variables. The tests are analogous to analysis of variance or t-tests where the response variable is not continuous -- it is recorded only as a discrete number or category (see Figure 4.1). When the response variable is **ordinal** (possible values can be ordered into a logical sequence, such as low, medium and high) the familiar Kruskal-Wallis test can be used. When the response variable is **nominal** (no ordering to the categories, such as with different species of organism), contingency tables can assess association. When both variables are ordinal, Kendall's tau can test for significance in association.

14.1 Recording Categorical Data

Categorical variables are those whose possible values are not along a continuous scale (such as concentration), but may take on only one of a discrete number of values classed into one of several categories. Examples of categorical variables used in water resources studies are: presence or absence of a benthic invertebrate, whether an organic compound is above or below the detection limit, soil type, land use group, and location variables such as aquifer unit, gaging station, etc. To easily inspect the relationship between two categorical variables, the data are recorded as a matrix of counts (Figure 14.1). The matrix is composed of two categorical variables, one assigned to the columns and one to the rows. Both variables will take on several possible values. The entries in the cells of the matrix are the number of observations O_{ij} which fall into the ith row and jth column of the matrix.

	Variable 2		
Variable 1	Group 1	Group 2	Group 3
Response 1	O_{11}	O_{12}	O_{13}
Response 2	O_{21}	O_{22}	O_{23}

Figure 14.1 Structure of a 2-variable matrix

14.2 Contingency Tables (Both Variables Nominal)

Contingency tables measure the association between two nominal categorical variables. Because they are nominal there is no natural ordering of either variable, so that categories may be switched in assignment from the first to the second row, etc. without any loss in meaning. The purpose of contingency table analysis is to determine whether the row classification (variable 1, here arbitrarily assigned to the response variable if there is one) is independent of the column classification (variable 2, here assigned to the location or group-of-origin variable). The null hypothesis H_O is that the two variables are independent -- that is, the distribution of data among the categories of the first variable is not affected by the classification of the second variable. Evidence may be sufficient to reject H_O in favor of H_1: the variables are dependent or related. The statement that one variable <u>causes</u> the observed values for the second variable is not necessarily implied, analogous to correlation. Causation must be determined by knowledge of the relevant processes, not only mathematical association. For example, both variables could be caused by a third underlying variable.

Example 1

Three streams are sampled to determine if they differ in their macrobiological community structure. In particular, the presence or absence of two species are recorded for each stream, one species being pollution tolerant, and one not. If the streams differ in their proportion of pollution-tolerant species, it is inferred that they differ in their pollution sources as well. Test whether the streams are identical in (independent of) the proportion of pollution-tolerant organisms, or whether they differ in this proportion (proportion is dependent on the stream).

H_O: the proportion of one species versus the second is the same for (is independent of) all 3 streams.

H_1: the proportion differs between (is dependent on) the stream.

	Stream 1	Stream 2	Stream 3	
Tolerant	O_{11}	O_{12}	O_{13}	$A_1 = \Sigma(O_{11}+O_{12}+O_{13})$
Intolerant	O_{21}	O_{22}	O_{23}	$A_2 = \Sigma(O_{21}+O_{22}+O_{23})$
	$C_1 = \Sigma(O_{11}+O_{21})$	$C_2 = \Sigma(O_{12}+O_{22})$	$C_3 = \Sigma(O_{13}+O_{23})$	**N = (A1+A2) = (C1+C2+C3)**

14.2.1 Performing the Test for Independence

To test for independence, the observed counts O_{ij} (row i and column j) in each cell are summed across rows to produce the row totals A_i, and down columns to produce column totals C_j. There are m rows (i=1,m) and k columns (j=1,k). The total sample size N is the sum of all counts in every cell, or $N = \Sigma A_i = \Sigma C_j = \Sigma O_{ij}$. The **marginal probability** of being in a given row (a_i) or column (c_j), can be computed by dividing the row A_i and column C_i totals by N:

$$a_i = A_i/N \qquad\qquad c_j = C_j/N$$

If H_O is true, the probability of a new sample falling into row 1 (species tolerant of pollution) is best estimated by the marginal probability a_1 regardless of which stream the sample was taken from. Thus the marginal probability for a row ignores any influence of the column variable.

The column variable is important in that the number of available samples may differ among the columns. The probability of being in any column may not be (1/no. columns). Therefore, with H_O true, the best estimate of the **joint probability** e_{ij} of being in a single cell in the table equals the marginal probability of being in row i times the marginal probability of being in column j

$$e_{ij} = a_i \bullet c_j.$$

Finally, for a sample size of N, the expected number of observations in each cell given H_O is true can be computed by multiplying each joint probability e_{ij} by N:

$$E_{ij} = N\, a_i\, c_j \;, \quad \text{or}$$

$$E_{ij} = \frac{A_i C_j}{N} \qquad [14.1]$$

To test H_O, a test statistic Xct is computed by directly comparing the observed counts O_{ij} with the counts E_{ij} expected when H_O is true. This statistic is the sum of the squared differences divided by the expected counts, summed over all i•j cells:

$$Xct = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad [14.2]$$

If H_O is not true, the observed counts O_{ij} will be very different from the E_{ij} for at least one cell and Xct will then be large. If H_O is true, the $O_{ij} \cong E_{ij}$ for all i•j cells, and Xct will be small. To evaluate whether Xct is sufficiently large to reject H_O, the test statistic is compared to the $(1-\alpha)$ quantile of a chi-square distribution having $(m-1)\bullet(k-1)$ degrees of freedom. Tables of the chi-square distribution are available in most statistics texts.

To understand why there are $(m-1)\bullet(k-1)$ degrees of freedom, when the marginal sums A_{ij} and C_{ij} are known, once $(m-1)\bullet(k-1)$ of the cell counts O_{ij} are specified the remainder can be computed. Therefore, only $(m-1)\bullet(k-1)$ entries can be "freely" specified.

<u>Example 1 cont.</u>
For the table of observed counts O_{ij} below, determine whether the streams differ significantly in their proportion of pollutant-tolerant species.

O_{ij}	Stream 1	Stream 2	Stream 3	
Tolerant	4	8	12	$A_1 = 24$
Intolerant	18	12	6	$A_2 = 36$
	$C_1=22$	$C_2=20$	$C_3=18$	**N = 60**

To determine whether the proportion of pollutant-tolerant species is significantly different for the three streams, a table of expected counts E_{ij} assuming H_O to be true is computed using equation 14.1:

E_{ij}	Stream 1	Stream 2	Stream 3	
Tolerant	8.8	8.0	7.2	$A_1 = 24$
Intolerant	13.2	12.0	10.8	$A_2 = 36$
	$C_1 = 22$	$C_2 = 20$	$C_3 = 18$	**60**

Dividing these expected counts by N results in the table of expected probabilities ($e_{ij} = E_{ij} / N$):

e_{ij}	Stream 1	Stream 2	Stream 3	
Tolerant	.148	.132	.120	$a_1 = 0.4$
Intolerant	.222	.198	.180	$a_2 = 0.6$
	$c_1 = 0.37$	$c_2 = 0.33$	$c_3 = 0.30$	**1.0**

To perform the significance test:

$$Xct = \frac{(4.0-8.8)^2}{8.8} + \frac{(8-8.0)^2}{8.0} + \frac{(12-7.2)^2}{7.2} + \frac{(18-13.2)^2}{13.2} + \frac{(12-12)^2}{12} + \frac{(6-10.8)^2}{10.8}$$

$$= 9.70$$

H_O should be rejected if Xct exceeds the $(1-\alpha)$ quantile of the chi-square distribution with $1 \bullet 2 = 2$ degrees of freedom. For $\alpha = 0.05$, $\chi^2_{(.95,\ 2)} = 5.99$. Therefore, H_O is rejected. The proportion of pollutant-tolerant species is not the same in all three streams. Thus the overall marginal probability of 0.4 is not an adequate estimate of the probability of pollution-tolerant species for all three streams.

14.2.2 Conditions Necessary for the Test

The chi-square distribution is a good approximation to the true distribution of Xct as long as

- all $E_{ij}>1$ and
- at least 80% of cells have $E_{ij} \geq 5$ (Conover, 1980).

If either condition is not met,

- combine two or more rows or columns and recompute, or
- enumerate the exact distribution of Xct. See Conover (1980) for details.

A contingency table test is not capable of extracting the information contained in any natural ordering of rows or columns. Contingency tables are designed to operate on nominal data without this ordering. The columns or rows can be rearranged without changing the expected values E_{ij}, and therefore without altering the test statistic. When the response variable or both variables have a natural scale of ordering, the test statistic should change as the ordinal variable is rearranged. Methods more powerful than contingency tables should be used when one or

both variables are ordinal. When only the response variable is ordinal, the Kruskal-Wallis test of the next section will have more power to see differences between groups than will contingency tables. When both variables are ordinal, Kendall's tau can measure the relationship as shown in section 14.4.

14.2.3 Location of the Differences

When a contingency table finds an association between the two variables, it is usually of interest to know how the two are related. Which cells are higher or lower in proportion than would be expected had H_O been true? A guide to this are the individual cell chi-square statistics.

Cells having large values of $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ are the cells contributing most to the rejection of the null hypothesis. The sign of the difference between O_{ij} and E_{ij} indicates the direction of the departure. For example, the individual cell chi-square statistics for the species data of example 1 are as follows:

χ^2	Stream 1	Stream 2	Stream 3
Tolerant	$\frac{(4.0-8.8)^2}{8.8} = 2.6$	$\frac{(8-8.0)^2}{8.0} = 0$	$\frac{(12-7.2)^2}{7.2} = 3.2$
Intolerant	$\frac{(18-13.2)^2}{13.2} = 1.7$	$\frac{(12-12)^2}{12} = 0$	$\frac{(6-10.8)^2}{10.8} = 2.1$

Stream 3 has many more counts of the pollution-tolerant species than the number expected if all three streams were alike, and stream 1 has many less. Therefore stream 1 appears to be the least affected by pollution, stream 2 in-between, and stream 3 the most affected.

14.3 Kruskal-Wallis Test for Ordered Categorical Responses

In Chapter 5 the Kruskal-Wallis test was introduced as a nonparametric test for differences in medians among 3 or more groups. The response variable in that case was continuous. In Chapter 13 the test was applied to response data whose lower end of a continuous scale was below a reporting limit. All censored data were treated as ties. Now the test will be applied to data which are ordinal -- the response variable can only be recorded as belonging to one of several ordered categories. All observations in the same response category (row) are tied with each other. The test takes on its most general form in this situation, as a test for whether a shift in the distribution has occurred, rather than as a test for differences in the median of continuous data. The test may be stated as:

H_O: the proportion of data in each response category (row) is the same for each group (column).

H_1: the proportion differs among (is dependent on) the groups.

14.3.1 Computing the Test

The data are organized in a matrix identical to that for a contingency table, but the computations at the margins differ (Figure 14.2). Once the row sums A_i are computed, ranks R_i are assigned to each observation in the table in accord with levels of the response variable. Ranks for all observations in the category with the lowest responses (response row 1 in Figure 14.2) will be tied at the average rank for that row, or $\overline{R}_1 = (A_1 + 1)/2$. All observations in the row having the next highest response are also assigned ranks tied at the average of ranks within that row, and so on up to the highest row of responses. For response 2 in the Figure 14.2 the average rank equals $\overline{R}_2 = A_1 + (A_2 + 1)/2$. For any row x of a total of m rows, the average rank will equal

$$\overline{R}_x = \sum_{i=1}^{x-1} A_i + (A_x+1)/2. \qquad [14.3]$$

To determine whether the distribution of proportions differs among the k groups (the k columns), the average column ranks $\overline{D}_j$ are computed as

$$\overline{D}_j = \frac{\sum_{i=1}^{m} O_{ij}\overline{R}_i}{C_j} \qquad \text{where } C_j = \sum_{i=1}^{m} O_{ij}. \qquad [14.4]$$

	Group 1	Group 2	Group 3	
response 1	O_{11}	O_{12}	O_{13}	$A_1 = \Sigma(O_{11}+O_{12}+O_{13})$
response 2	O_{21}	O_{22}	O_{23}	$A_2 = \Sigma(O_{21}+O_{22}+O_{23})$
	$\overline{D}_1$	$\overline{D}_2$	$\overline{D}_3$	**N**

where

$$\overline{D}_1 = \frac{(O_{11}\overline{R}_1 + O_{21}\overline{R}_2)}{O_{11}+O_{21}} \qquad \overline{D}_2 = \frac{(O_{12}\overline{R}_1 + O_{22}\overline{R}_2)}{O_{12}+O_{22}} \qquad \overline{D}_3 = \frac{(O_{13}\overline{R}_1 + O_{23}\overline{R}_2)}{O_{13}+O_{23}}$$

Figure 14.2 2x3 matrix for Kruskal-Wallis analysis of an ordered response variable

The Kruskal-Wallis test statistic is then computed from these average group ranks. If H_O is true, the average ranks $\overline{D}_j$ will all be about the same, and similar to the overall average rank of (N+1)/2. If H_O is not true, the average rank for at least one of the columns will differ. The Kruskal-Wallis test statistic is computed using equation 14.5:

$$K = (N-1)\ \frac{\sum_{j=1}^{k}(C_j \overline{D}_j^2) - N\left[\frac{N+1}{2}\right]^2}{\sum_{i=1}^{m}(A_i \overline{R}_i^2) - N\left[\frac{N+1}{2}\right]^2} \qquad [14.5]$$

where C_j is the number of observations in column j,
$\overline{D}_j$ is the average rank of observations in column j,
A_i is the number of observations in row i, and
$\overline{R}_i$ is the average rank of observations in row i.

To evaluate its significance, K is compared to a table of the chi-square distribution with k−1 degrees of freedom.

<u>Example 2</u>
An organic chemical is measured in 60 wells screened in one of three aquifers. The concentration is recorded only as being either above or below the reporting limit (rl). Does the distribution of the chemical differ among the three aquifers?

First, ranks are assigned to the response variable. There are 36 observations in the lower category (below rl), each with a rank equal to the mean rank of that group. The mean of numbers 1 to 36 is (36+1)/2 = 18.5. The higher category contains 24 observations with ranks 37 to 60, so that their mean rank is 36 + (24+1)/2, or 48.5.

	Aquifer 1	Aquifer 2	Aquifer 3	A_i	$\overline{R}_i$
below rl	18	12	6	36	18.5
above rl	4	8	12	24	48.5
	$\overline{D}_1$ = 527/22 = 24	$\overline{D}_2$ = 610/20 = 30.5	$\overline{D}_3$ = 693/18 = 38.5		

$$K = (59)\ \frac{\sum (22 \bullet 24^2 + 20 \bullet 30.5^2 + 18 \bullet 38.5^2) - 60\left[\frac{61}{2}\right]^2}{\sum (24 \bullet 48.5^2 + 36 \bullet 18.5^2) - 60\left[\frac{61}{2}\right]^2}$$

$$= 9.75$$

The chi-square statistic $\chi^2_{(.95,\ 2)} = 5.99$. Thus H_O is rejected, and the groups are found to have differing percentages of data above the reporting limit.

14.3.2 Multiple Comparisons

Once differences between the groups (columns) have been found, it is usually of interest to determine which groups differ from others. This is done with multiple comparison tests as stated in section 7.4. Briefly, multiple Kruskal-Wallis tests are performed between pairs of columns. After a significant KW test occurs for k groups, place the groups in order of ascending average rank. Perform new KW tests for the two possible comparisons between groups which are p = (k–1) columns apart (the first versus the next-to-last column, and the second versus the last). Note that the observations must be re-ranked for each test. If significant results occur for one or both of these tests, continue attempting to find differences between smaller subsets of any groups found to be significantly different. In order to control the overall error rate, set the individual error rates for each KW test at α_p, below:

$$\alpha_p = 1 - (1-\alpha)^{p/k} \qquad \text{for } p < (k-1)$$
$$= \alpha \qquad \text{for } p \geq (k-1)$$

14.4 Kendall's Tau for Categorical Data (Both Variables Ordinal)

When both row and column variables are ordinal, a contingency table would test for differences in distribution of the row categories among the columns, but would ignore the correlation structure of the data -- do increases in the column variable coincide with increases or decreases in the row variable? This additional information contained in the correlation structure of ordinal variables can be evaluated with a rank correlation test such as Kendall's tau.

14.4.1 Kendall's τ_b for Tied Data

Kendall's correlation coefficient tau (τ) must be modified in the presence of ties. In Chapter 8 a tie modification was given for ties in the response variable only. Now there are many more ties, the ties between all data found in the same row and column of a contingency table. Kendall (1975) called this tie modification τ_b (tau-b).

$$\tau_b = \frac{S}{\frac{1}{2}\sqrt{(N^2 - SS_a)\,(N^2 - SS_c)}} \qquad [14.6]$$

The numerator S for τ_b is P–M, just as in Chapter 8, the number of pluses minus the number of minuses. Consider a contingency table structure with the lowest values on the upper left (the

rows are ordered from lowest value on the top to the highest value on the bottom, and the columns from lowest on the left to highest on the right -- see Figure 14.3). With this format, the number of pluses are the number of comparisons with data in cells to the right and below each cell (Figure 14.4). The number of minuses are the number of comparisons with data in cells to the left and below (Figure 14.5). Data in cells of the same row or column do not contribute to S. Therefore, summing over each cell of row x and column y,

$$S = P - M = \sum_{xy} O_{xy} \left(\sum O_{southeast} - \sum O_{southwest}\right), \text{ or}$$

$$S = \sum_{\text{all } x\ y} \sum_{i>x} \sum_{j>y} O_{xy} \bullet O_{ij} - \sum_{i<x} \sum_{j<y} O_{xy} \bullet O_{ij} \qquad [14.7]$$

The denominator for τ_b is not $(n \bullet n-1)/2$ as it was for τ, equal to the total number of comparisons. Instead S is divided by the total number of untied comparisons. To compute this efficiently with a contingency table, SS_a and SS_c (the sums of squares of the row and column marginal totals, respectively) are computed as in equation 14.8, and then used in equation 14.6 to compute τ_b.

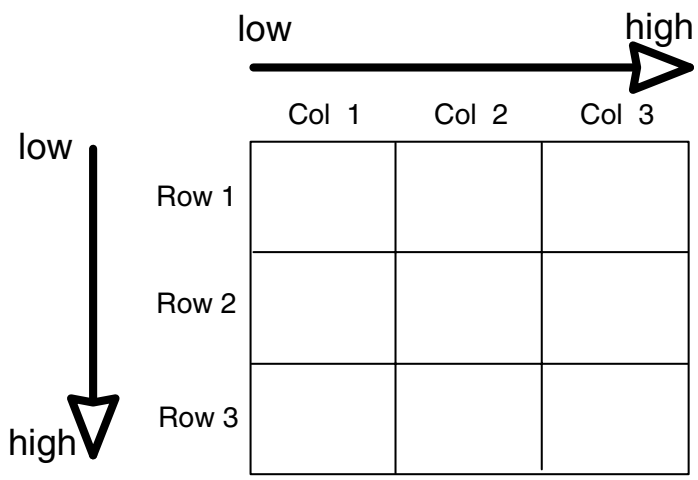


Figure 14.3 Suggested ordering of rows and columns for computing τ_b.

$$SS_a = \sum_{i=1}^{m} A_i^2 \qquad SS_c = \sum_{j=1}^{k} C_j^2 \qquad [14.8]$$

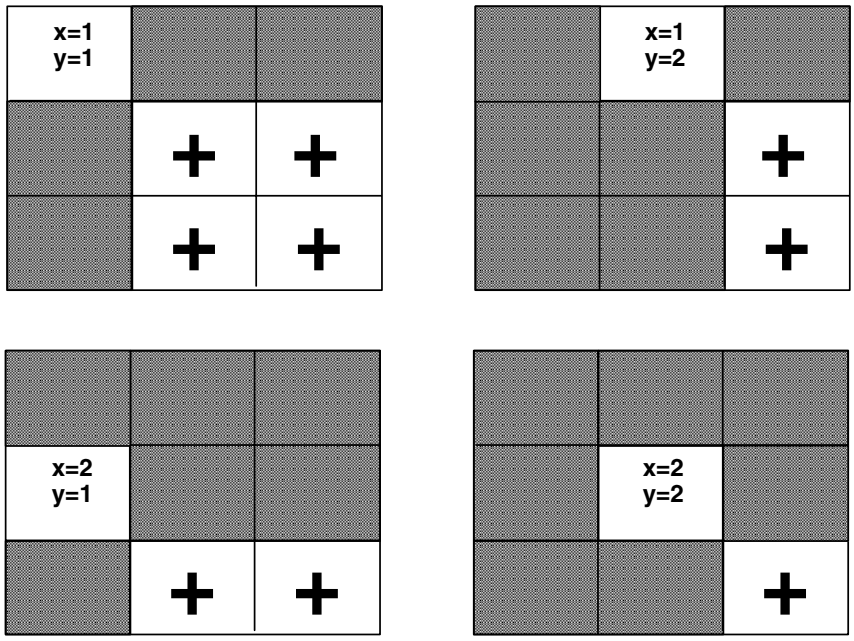


Figure 14.4 3x3 matrix cells contributing to P (i>x and j>y).

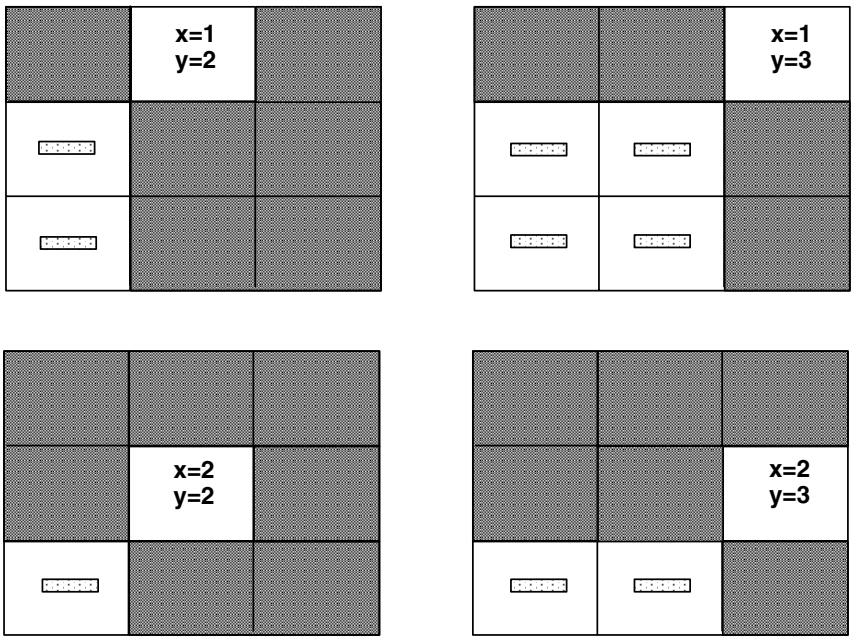


Figure 14.5 3x3 matrix cells contributing to M (i<x and j<y).

<u>Example 3</u>

Pesticide concentrations in shallow aquifers were investigated to test whether their distribution was the same for wells located in three soil classes, or whether concentrations differed with increasing soil drainage. The laboratory reported concentrations for the pesticide when levels

were above the reporting limit. The compound was reported only as "present" when concentrations were between the reporting limit and the instrument detection limit (dl), and as "<dl" if concentrations were below the detection limit. Compute Kendall's tau for this data.

	Soil Drainage				
Concentration	Poor	Moderate	High	A_i	a_i
< dl	18	12	7	37	0.47
"present"	5	10	8	23	0.29
> rl	2	6	11	19	0.24
C_j	25	28	26	**79**	
c_j	0.32	0.35	0.33		**1.0**

The number of pluses P = 18(10+8+6+11) + 12(8+11) + 5(6+11) + 10(11) = 1053
The number of minuses M = 12(5+2) + 7(5+10+2+6) + 10(2) + 8(2+6) = 329
So S = 1053 – 322 = 731.

To compute the denominator of τ_b, $SS_a = 37^2 + 23^2 + 19^2 = 2259.$
$SS_c = 25^2 + 28^2 + 26^2 = 2085.$

and $$\tau_b = \frac{724}{\frac{\sqrt{(79^2 - 2259)(79^2 - 2085)}}{2}} = \frac{724}{2034} = 0.36.$$

14.4.2 Test of Significance for τ_b

To determine whether τ_b is significantly different from zero, S must be divided by its standard error σ_S and compared to a table of the normal distribution, just as in Chapter 8. Agresti (1984) provides the following approximation to σ_S which is valid for P and M larger than 100:

$$\sigma_S \cong \sqrt{\frac{1}{9} \cdot \left(1 - \sum_{i=1}^{m} a_i^3\right)\left(1 - \sum_{j=1}^{k} c_j^3\right) \cdot N^3} \qquad [14.9]$$

where a_i and c_j are the marginal probabilities of each row and column.

The exact formula for σ_S (Kendall, 1975) is much more complicated. It is the square root of equation 14.10:

$$\sigma_S^2 = \frac{\left(n(n-1)(2n+5) - \sum_{i=1}^{m} A_i(A_i-1)\,(2A_i+5) - \sum_{j=1}^{k} C_j(C_j-1)\,(2C_j+5)\right)}{18} +$$

$$\frac{\left(\sum_{i=1}^{m} A_i(A_i-1)(A_i-2)\right)\left(\sum_{j=1}^{k} C_j(C_j-1)(C_j-2)\right)}{9 \bullet N(N-1)(N-2)} + \frac{\left(\sum_{i=1}^{m} A_i(A_i-1)\right)\left(\sum_{j=1}^{k} C_j(C_j-1)\right)}{2 \bullet N(N-1)} \qquad [14.10]$$

If one variable were continuous and contained no ties, equation 14.10 would simplify to the square of equation 8.4.

To test for significance of τ_b, the test statistic Z_S is computed as in Chapter 8:

$$Z_S = \begin{cases} \dfrac{S-1}{\sigma_s} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \dfrac{S+1}{\sigma_s} & \text{if } S < 0 \end{cases} \qquad [14.11]$$

Z_S is compared to the $\alpha/2$ quantile of the normal distribution to obtain the two-sided p-value for the test of significance on τ_b.

<u>Example 3, cont.</u>
Is the value of $\tau_b = 0.36$ significantly different from zero? From equation 14.9 the approximate value of σ_S is

$$\sigma_S \cong \sqrt{\frac{1}{9} \bullet \left(1-(0.47^3+0.29^3+0.24^3)\right) \bullet \left(1-(0.32^3+0.35^3+0.33^3)\right) \bullet 79^3}$$

$$\cong \sqrt{\frac{(0.86)\bullet(0.89)\bullet 79^3}{9}} = \sqrt{42329} = 205.74$$

$$Z_S \cong \frac{724-1}{205.74} = 3.51$$

and from a table of the normal distribution the one-sided p-value is $p = 0.0002$. Therefore H_O: $\tau_b = 0$ is rejected, which means that pesticide concentrations increase (the distribution shifts to a greater proportion of higher classes) as soil drainage increases.

14.5 Other Methods for Analysis of Categorical Data

One other method is prominent in the statistical literature for analysis of all three situations discussed in this chapter -- loglinear models (Agresti, 1984). Loglinear models transform the expected cell probabilities $e_{ij} = a_i \bullet c_j$ by taking logarithms to produce a linear equation $\ln(e_{ij}) = \mu + \ln(a_i) + \ln(c_j)$, where μ is a constant. Models may be formulated for the completely nominal case, as well as for one or more ordinal variables. Detailed contrasts of the probability of being in column 2 versus column 1, column 3 versus 2, etc. are possible using the loglinear model. Tests for higher dimensioned matrices (such as a 3-variable 3x2x4 matrix) can be formulated. Interactions between the variables may be formulated and tested analogous to an analysis of variance on continuous variables. Though the computation of such models is not discussed here, Agresti (1984) provides ample detail.

Exercises

14.1 Samples of water quality collected at USGS National Stream Quality Accounting Network (NASQAN) stations from 1974 to 1981 show more frequent increases in chloride ion than decreases. 245 stations were classified by Smith et al. (1987) by their trend analysis results at $\alpha = 0.1$. One reasonable cause for observed trends is road salt applications. Estimates of tons of road salt applied to the 245 basins in 1975 and 1980 are used to place the stations into into 3 groups: decreases (1980 is more than 20% less than 1975), increases (1980 is more than 20% greater than 1975), and little or no change. The two variables are then summarized by this 3x3 table:

	Trend in Cl^- (1974-81, α=0.1)			
Δ road salt appl.	Down	No trend	Up	Totals
Down	5	32	9	46
No change	14	44	25	83
Up	10	51	55	116
Totals	29	127	89	**245**

Test H_O: a basin's trend in chloride ion is independent of its change in road salt application, versus the alternative that they are related.

a) using a contingency table. Interpret the test result.
b) using Kendall's tau. Interpret the test result.
c) which test is more appropriate, and why?

14.2 Fusillo et al. (1985) sampled 294 wells in New Jersey for volatile organic compounds. The wells were classified by whether they were located in an outcrop location near the surface, or whether they were further downdip and somewhat more protected from direct contamination. Determine whether the probability of finding detectable levels of volatile compounds differs based on the location of the well.

Location	Non-detects	Detect VOC	Totals
Downdip	106	9	115
Outcrop	129	50	179
Totals	235	59	**294**

14.3 Regulation of organo-tin antifouling paints for boats was announced in 1988 in Switzerland. Concentrations of tributyltin (TBT, in ng/L) in unfiltered water samples from Swiss marinas were measured in 1988 to 1990 (Fent and Hunn, 1991). Is there evidence of a decrease in TBT concentrations in marina waters as these paints were being taken off the market?

	Number of samples		
Year	TBT≤ 200	TBT>200	Totals
1988	2	7	
1989	9	13	
1990	10	10	
Totals			**51**

Chapter 15

Regression for Discrete Responses

Concentrations of a volatile organic chemical are measured in numerous wells across a large study area. About 75% of the resulting samples are below the laboratory reporting limit. The likelihood of finding concentrations above this limit is suspected to be a function of several variables, including population density, industrial activity, and traffic density. What is the most appropriate way to model the probability of being above the reporting limit using a regression-like relationship?

Streams can be classified according to whether or not they meet some criteria for use set by a regulatory agency. For example, a stream may be considered "fishable" or "not fishable", depending on several concentration and esthetic standards. What is the probability that a stream reach will meet the "fishable" criteria as a function of population density, distance downstream from the nearest point source, and percentage of the basin used for crop agriculture?

The above situations involve fitting a model similar to OLS regression, in that the explanatory variables are continuous. However the response variable is discrete -- it can be designated by an integer value (see figure 4.1). Discrete (or categorical) response variables are often encountered when the measurement process is not sufficiently precise to provide a continuous scale. Instead of an estimate of concentration, for example, only whether or not a sample exceeds some threshold, such as a reporting limit or health standard, is recorded. In water resources this response is usually ordinal. Logistic regression is the most commonly-used procedure for this situation. The equation predicts the probability of being in one of the possible response groups.

Discrete response variables are commonly binary (two categories). For example, species of organism or attribute of an organism are listed as either present or absent. Analysis of binary responses using logistic regression is discussed in the following sections, beginning with 15.1. Analysis of multiple response categories is discussed in section 15.4.

15.1 Regression for Binary Response Variables

With OLS regression, the actual magnitude of a response variable is modelled as a function of the magnitudes of one or more continuous explanatory variables. When the response is a binary categorical variable, however, it is the probability p of being in one of the two response groups that is modelled. The response variable is coded by setting the larger of the two possible responses (above or present) equal to 1, and the lower to 0. The predicted probability p is then the probability of the response being a 1, with 1−p as the probability of the response being a 0. The explanatory variables may be either continuous as in OLS regression, or a mixture of continuous and discrete variables similar to analysis of covariance. If all explanatory variables are discrete, logistic regression provides a multivariate alternative to the test for significance by Kendall's tau used in Chapter 14.

15.1.1 Use of Ordinary Least Squares

In the case of a binary response, the attempt to predict $\hat{p}$ = the probability of a response of 1 could be done with OLS regression. This would be a simple but incorrect approach. There are three reasons why this is not appropriate (Judge, et al., 1985):

1. Predictions $\hat{p}$ may fall outside of the 0 to 1 boundary.
2. The variance of $\hat{p}$ is not constant over the range of x's, violating one of the basic assumptions of OLS. Instead, the variance of the binary response variable equals $p \bullet (1-p)$, where p is the true probability of a 1 response for that x. Because this is not constant over x, weighted least squares must be used to obtain minimum variance and unbiased estimates of slope and intercept. See Draper and Smith (1981, pp. 108-116) for the WLS approach. WLS is still not appropriate, however, if estimates go beyond the 0 to 1 boundary.
3. Residuals from the regression cannot be normally distributed. This renders tests on the slope coefficients invalid.

OLS been used with discrete responses when multiple observations occur for all or most combinations of explanatory variables. The responses (0 or 1) are first grouped by some range of explanatory variable(s). This creates a new continuous y variable, the proportion of responses which equal 1. Even so, least-squares regression fails the three criteria above, so that more appropriate methods are warranted.

15.2 Logistic Regression

Logistic regression, also called logit regression, transforms the estimated probabilities $\hat{p}$ into a continuous response variable with values possible from – to + infinity. The transformed response is predicted from one or more explanatory variables, and subsequently retransformed back to a value between 0 and 1. A plot of estimated probabilities has an S shape (figure 15.1). The estimates of probability change most rapidly at the center of the data. Thus logistic regression is most applicable for phenomena which change less rapidly as p approaches its limits of 0 or 1. However, when the range of predicted probabilities does not get near its extremes, the plot is one of mild curvature (figure 15.2). Thus the function is a flexible and useful one for many situations. A review of this and other categorical response models is given by Amemiya (1981).

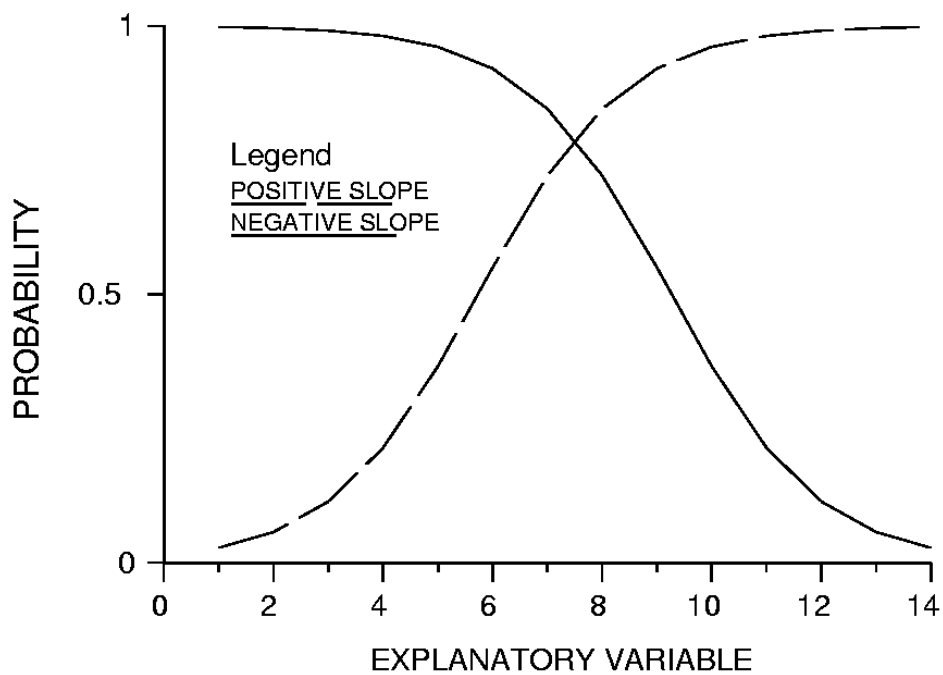


Figure 15.1 Logistic regression equations with – and + slopes. Note that estimates change more rapidly in the center than at the extremes.

15.2.1 Important Formulae

The **odds ratio** is defined as the ratio of the probability of obtaining a 1 divided by the probability of obtaining a 0:

$$\text{odds ratio} = \left(\frac{p}{1 - p}\right) \qquad [15.1]$$

where p is the probability of a response of 1.

The log of the odds ratio or **logit** transforms a variable constrained between 0 and 1, such as a proportion, into a continuous and unbounded variable. The logit can then be modeled as a linear function of one or more explanatory variables to produce logistic regression:

$$\log\left(\frac{p}{1 - p}\right) = b_0 + \mathbf{bX} \qquad [15.2]$$

where b_0 is the intercept, **X** is a vector of k explanatory variable(s), and **bX** includes the slope coefficients for each explanatory variable so that $\mathbf{bX} = b_1X_1, b_2X_2, \ldots b_kX_k$.

Thus the odds ratio is modelled as

$$\left(\frac{p}{1 - p}\right) = \exp(b_0 + \mathbf{bX}). \qquad [15.3]$$

To return the predicted values of the response variable to original units, the logistic transformation (the inverse of the logit transformation) is used:

$$p = \frac{\exp(b_0 + \mathbf{bX})}{[1 + \exp(b_0 + \mathbf{bX})]} \qquad [15.4]$$

For example, the mutliple logistic regression equation with two explanatory variables would look like

$$p = \frac{\exp(b_0 + b_1X_1 + b_2X_2)}{[1 + \exp(b_0 + b_1X_1 + b_2X_2)]}$$

For a single x variable, the odds of obtaining a 1 response increase multiplicatively by e^{b_1} for every unit increase in X. The inflection point of the curve is at $-b_0/b_1$, which is the median of the data. The slope of the estimated probability is greatest at this point. Equations are analogous for multiple explanatory variables. Biologists call the inflection point the median lethal dose (LD_{50}) when predicting the probability of death from some concentration (dose) of toxicant. The animal has a 50% chance of survival at this dose.

15.2.2 Computation by Maximum Likelihood

Estimates b_j of the j=1,...k slope coefficients could physically be computed by WLS when the input data are proportions between 0 and 1 (but they should not -- see section 15.1.1). However, the original data are most often coded only in the binary form, with replicates not available for computing proportions. A more general method for computing slope coefficients, valid for both binary and proportions as input data, is maximum likelihood estimation. Maximum likelihood optimizes the likelihood that the observed data would be produced from a given set of slopes. It is an iterative procedure available in the more complex statistical software packages. A function called the **log likelihood** (l) of the overall regression model is written as:

$$l = \sum_{i=1}^{n} \left(y_i \bullet \ln[\hat{p}_i] + (1-y_i) \bullet \ln[1-\hat{p}_i] \right) \qquad [15.5]$$

for the i=1,n binary observations y_i and predicted probabilities $\hat{p}$. When $y_i = 0$, the second term inside the brackets is nonzero, and a $\hat{p}$ is desirable which is close to 0. When $y_i = 1$, the first term is nonzero and a $\hat{p}$ close to 1 is desirable. The log of either $\hat{p}$ or $[1-\hat{p}]$ will be negative, and therefore l is a negative number which is maximized (brought closest to 0) by iteratively substituting estimates of p derived from estimates of slopes and intercept. The log likelihood may be alternately reported as the positive number G^2, the **−2 log likelihood**, which is minimized by the MLE procedure:

$$-2 \text{ log likelihood } G^2 = -2 \bullet l. \qquad [15.6]$$

15.2.3 Hypothesis Tests

15.2.3.1 Test for overall significance

An overall test of whether a logistic regression model fits the observed data better than an intercept-only model (where all slopes $b_j = 0$), analogous to the overall F test in multiple regression, is given by the **overall likelihood ratio** (lr_o):

$$lr_o = 2 \bullet (l - l_0) = (G^2_0 - G^2) \qquad [15.7]$$

where l is the log likelihood of the full model, l_0 is the log likelihood of the intercept-only model, and G^2_0 is the −2 log likelihood of the intercept only model.

The overall likelihood ratio lr_o can be approximated by a chi-square distribution with k degrees of freedom, where k is the number of slopes estimated. If $lr_o > \chi^2_{k,\alpha}$ then the null hypothesis that all $b_j = 0$ can be rejected. Should the null hypothesis not be rejected, the best estimate over all **X** of the probability of a 1 is simply the proportion of the entire data set which equals 1.

15.2.3.2 Testing nested models

To compare nested logistic regression models, similar to the partial F tests in OLS regression, the test statistic is the **partial likelihood ratio** lr:

$$lr = 2 \bullet (l_c - l_s) = (G^2_s - G^2_c) \qquad [15.8]$$

where l_c is the log likelihood for the more complex model, and
l_s is the log likelihood for the simpler model.

The partial likelihood ratio is approximated by a chi-square distribution with $(k_c - k_s)$ degrees of freedom, the number of additional coefficients in the more complex model. For the case where only one additional coefficient is added, the chi-square with 1 degree of freedom equals the

square of a t-statistic called **Wald's t**, computed from the estimated coefficient b divided by its standard error. Degrees of freedom for the t-statistic are the number of observations n minus the number of estimated slopes, or n–k. As with OLS regression, some computer software will report the t-statistic, while others report the $t^2 = \chi^2$ value; p-values will be essentially the same for either form of the test.

15.2.4 Amount of Uncertainty Explained, R^2

A measure of the amount of uncertainty explained by the model, actually the proportion of log-likelihood explained, is McFadden's R^2, or the **likelihood-R^2**,

$$R^2 = 1 - \frac{l}{l_0} \qquad [15.9]$$

where l and l_0 are as before. The likelihood-R^2 is uncorrected for the number of coefficients in the model. much like R^2 in OLS regression.

A second measure of the amount of uncertainty explained by the model is the R^2 between the observed and predicted values of p, or **Efron's R^2**

$$\text{Efron's } R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{p})^2}{\sum_{i=1}^{n} (y_i - \bar{p})^2} \qquad [15.10]$$

where $\bar{p} = \sum y_i/n$, the proportion = 1 for the entire data set. However, this version of R^2 is not as appropriate as the likelihood-R^2 because the residuals $(y_i - \hat{p})$ are heteroscedastic due to the binary nature of the y_i.

15.2.5 Comparing Non-Nested Models

To compare two or more non-nested logistic regression models, partial likelihood ratios are not appropriate. This is the situation in OLS regression where Mallow's Cp or PRESS is used. For likelihood ratio tests, a statistic related to Mallow's Cp is **Akaike's Information Criteria (AIC)**. AIC includes both a measure of model error (–l) and a penalty for too many variables, the number of explanatory variables k. Better models are those with small AIC. Akaike's information criteria

$$AIC = -l + k \qquad [15.11]$$

AIC can also be written to expressly include the comparison of each candidate model to the full model (the model which includes all possible explanatory variables).

$$\begin{aligned} AIC^* &= 2\,(l_f - l) - 2\bullet(k_f - k) \\ &= (G^2 - G^2_f) - 2\bullet\Delta df \\ &= lr - 2\bullet\Delta df \end{aligned}$$

where l_f is the log likelihood of the full model, k_f is the degrees of freedom of the full model, Δdf is the difference in the degrees of freedom between the model and the full model, and lr is the partial likelihood ratio comparing the candidate and full models. Either form should be minimized to find the best model.

Related to the AIC is an **adjusted R^2** which adjusts for the degrees of freedom in the model. It penalizes a model which includes too many slope parameters. The adjusted R^2 allows comparisons between models with differing number of explanatory variables:

$$\text{adjusted } R^2 = 1 - \frac{(l - k)}{l_0} = 1 - \frac{2\bullet AIC}{G^2_0} \qquad [15.12]$$

This adjusted R^2 should be maximized.

Example 1

Eckhardt et al. (1989) reported the pattern of occurrence for several volatile organic compounds in shallow groundwaters on Long Island, NY. TCE detections for 643 samples are listed in table 15.1 below, where 1 signifies a concentration above the reporting limit of 3 ppb. Logistic regression between occurrence (1) or non-occurrence (0) as a function of population density gives the following results:

Population Density	no. 1s	no. 0s	N	%1s
1	1	148	149	0.7
2	4	80	84	4.8
3	10	88	98	10.2
5	25	86	111	22.5
6	11	33	44	25.0
8	8	24	32	25.0
9	29	14	43	67.4
11	19	31	50	38.0
13	6	5	11	54.5
14	2	11	13	15.4
17	2	5	7	28.6
19	0	1	1	0.0
overall	117	526	643	18.2

Table 15.1 TCE data in the Upper Glacial Aquifer, Long Island

The log likelihood for the intercept-only model $l_0 = -305.0$ ($G^2_0 = 610.0$). To determine the significance of population density (POPDEN) as an explanatory variable, the likelihood ratio is

computed by subtracting the log likelihood of this one-variable model from that of intercept-only model, and comparing to a chi-square distribution:

lr = 610.0 – 533.0 = 77.0 with 1 df resulting in a p-value = 0.0001.

Table 15.2 gives the important statistics for the model. A plot of the logistic regression line along with bars of ± 2 standard errors are shown in figure 15.2.

–2 log likelihood = 533.0

Explanatory variable	Estimate	Partial t-statistic	p-value
INTERCEPT	–2.80	–13.4	0.0001
POPDEN	0.226	8.33	0.0001

Table 15.2 Statistics for the popden model.

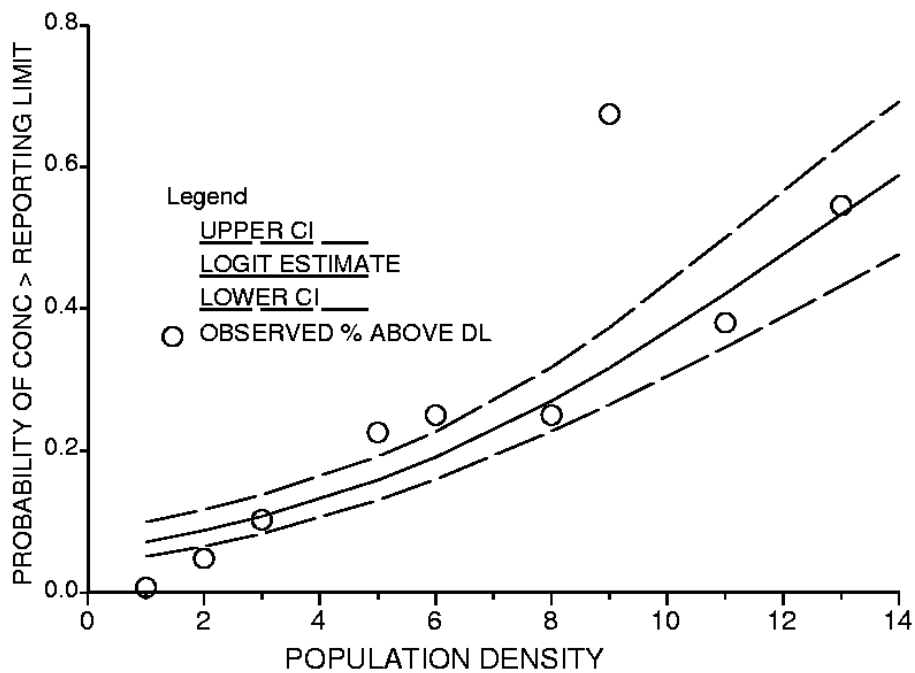


Figure 15.2 Logistic regression line for the TCE data, with percent detections observed for each population density.

The positive slope coefficient for popden means that the probability of a response =1 (concentration above the reporting limit) increases with increasing population density. Note that the line did not fit the observed data well at popden= 9. A second variable, a binary indicator of whether or not the area around the well was sewered, was added to the model in hopes of improving the fit. Does this second variable help explain more of the variation observed? The results are presented in table 15.3.

−2 log likelihood = 506.3

Explanatory variable	Estimate	Partial t-statistic	p-value
INTERCEPT	−3.24	−12.47	0.0001
POPDEN	0.13	4.07	0.0001
SEWER	1.54	4.94	0.0001

Table 15.3 Statistics for the popden + sewer model.

The likelihood ratio test determines whether this model is better than an intercept-only model

$lr_O = 610.0 - 506.3 = 103.7$ with 2 df resulting in a p-value = 0.0001.

Thus this logistic regression is significantly better than just estimating the proportion of data above the detection limit without regard to the two variables. The positive slope estimate for sewer means that the probability of detection of TCE increases with increasing proportion of sewering around the well. Note that this does not prove that sewering itself is the cause -- this could result from sewering as a surrogate for increasing urbanization or industrialization of the area. The usefulness of sewer in comparison to the popden-only model is seen by the significance of its partial t-statistic. It may also be measured by the difference in likelihood ratios for the one and two-variable models:

$lr = 533.0 - 506.3 = 26.7$ with 1 df resulting in a p-value = 0.0001.

Next a model with completely different explanatory variables was tried, relating TCE detections to the amount of land near the well which was classified as industrial land (indlu), and to the depth of the water below land surface. The results are given in table 15.4. As the partial t-statistics are both significant, a logical question is which of the two 2-variable models is preferable?

−2 log likelihood = 557.8

Explanatory variable	Estimate	Partial t-statistic	p-value
INTERCEPT	−1.07	−5.49	0.0001
INDLU	0.092	4.61	0.0001
DEPTH	0.008	−4.52	0.0001

Table 15.4 Statistics for the indlu + depth model.

As these models are not nested, they must be compared using AIC. Magnitudes of their partial t-statistics will not help decide which to use. As seen is table 15.5, the AIC for the population+sewer model is lower, and therefore is the preferable model between these two candidates.

Explanatory variables	–l	k (# exp. vars.)	AIC
POPDEN, SEWER	253.2	2	255.2
INDLU, DEPTH	278.9	2	280.9

Table 15.5 AIC for comparing two 2-variable logit models.

15.3 Alternatives to Logistic Regression

Two other methods have been used to relate one or more continuous variables to a binary variable.-- discriminant function analysis (parametric), and the nonparametric rank-sum test. In the following sections these methods are compared to logistic regression.

15.3.1 Discriminant Function Analysis

Discriminant function analysis is used as a multivariate classification tool, to decide in which of several groups a response is most likely to belong (Johnson and Wichern, 1982). Probabilities of being in each of the groups is computed as a function of one or more continuous variables. The group having the highest probability is selected as the group most likely to contain that observation. An equation (the discriminant function) is computed from data classified into known groups, and used to classify additional observations whose group affiliation is unknown. As each group is assigned an integer value, these objectives are identical to those of logistic regression.

The primary drawback of discriminant analysis is that it makes two assumptions:
1) multivariate normality, and 2) that the variance of data within each group is identical for all groups. Thus it requires the same assumptions as does a t-test or analysis of variance, but in multiple dimensions when multiple explanatory variables are employed. It will be slightly more efficient than logistic regression if these assumptions are true, but is much less robust (Press and Wilson, 1978). Therefore logistic regression should be preferred when multivariate normality and equality of variances cannot be assumed, as is the case for most of the data found in water resources.

15.3.2 Rank-Sum Test

Dietz (1985) has shown that the rank-sum test is a powerful alternative to the more complicated likelihood-ratio test for determining whether a binary response variable is significantly related to one continuous explanatory variable. The responses of 0 and 1 are treated as two separate groups, and the ranks of the continuous variable are tested for differences among the two response groups. When the probabilities of a 0 or 1 differ as a function of x, the ranks of x will differ between the two response variable groups. A slight modification to the rank-sum test is necessary for small sample sizes (see Dietz, 1985). The rank-sum test is equivalent to the significance test for Kendall's tau between the binary y variable and a continuous x.

When software is not available to perform likelihood-ratio tests, the rank-sum test can be used with little loss in power. However, it only considers the influence of one explanatory variable. There also is no slope estimate or equation associated with the rank-sum test when the responses are recorded as 0 or 1. When the responses are proportions between 0 and 1, Kendall's robust line may be used to linearly relate logits to the explanatory variable, though estimates below 0 or above 1 may result.

15.4 Logistic Regression for More Than Two Response Categories

In water resources applications, response variables may often be discretized into more than two response categories. Extensions of logistic regression for binary responses are available to analyze these situations. The method of analysis should differ depending on whether the response variable is ordinal or simply nominal. Ordinal responses such as low, medium and high are the most common situation in water resources. Here a common logit slope is computed, with multiple thresholds differing by offset intercepts in logit units. When responses are not ordinal, the possible response contrasts -- such as the probabilities of being in group 1 versus group 2 and in group 2 versus 3 -- are independent. In this case independent logit models are fit for each threshold.

15.4.1 Ordered Response Categories

Categorical response variables sometimes represent an underlying continuous variable which cannot be measured with precision sufficient to provide a continuous scale. For example, concentration data may be discretized into above and below a detection limit, or into three categories based on two thresholds (see below). Biologic activity may be categorized as not affected, slightly affected or severely affected by pollution. The resulting multiple responses y_i, i=1 to m are ordinal, so that $y_1<y_2<...<y_m$.

For example, suppose 3 responses are possible:

0: concentrations are below the reporting limit,
1: concentrations are above the reporting limit but below a health standard, and
2: concentrations are above the health standard.

This corresponds to two thresholds, one below versus above the reporting limit (0 versus not 0) and the second below versus above the health standard (not 2 versus 2). Figure 15.3 shows that for y=2, a transformation of the underlying continuous concentration Y* can be developed such that y=2 only when X>Y* for one explanatory variable X. Similarly, y>0 (above the reporting limit) only when $X>Y^*-\delta$, where δ is the difference between the two thresholds in the transformed scale. Therefore the upper threshold can be modeled as:

$$\log\left(\frac{\text{Prob}(y=2)}{\text{Prob}(y=0)+\text{Prob}(y=0)}\right) = \text{Prob}(X > Y^*) = b_0 + b_1 X\,, \qquad [15.13]$$

where b_0 is the estimate of intercept and b_1 the estimate of slope. This is a standard logistic regression identical to the binary case of not 2 versus 2. The probability of being above the lower threshold (reporting limit) is modelled using

$$\begin{aligned}\log\left(\frac{\text{Prob}(y=1)+\text{Prob}(y=2)}{\text{Prob}(y=0)}\right) = \text{Prob}(X > Y^*-\delta) &= b_0 + b_1(X+\delta), \qquad [15.14]\\ &= b_0' + b_1 X\\ &= b_0 + \lambda + b_1 X\end{aligned}$$

where $\lambda = b_1\delta$ is a shift parameter that must be estimated (McCullagh, 1980). Because the responses are ordered, the slope b_1 is common to all thresholds, and represents the proportional effect of X on the underlying and unobserved Y*. The resulting s-shaped curves for each threshold are simply offset (figure 15.4). Unfortunately the method for efficiently estimating these parameters is not available on many commercial statistics packages. McCullagh (1980) discusses the mathematics. As an alternative, separate logistic regressions can be estimated for each threshold (see below). This procedure is less efficient for the case of ordered responses, being appropriate for nominal responses. Unfortunately, it is the best that is available to most practitioners.

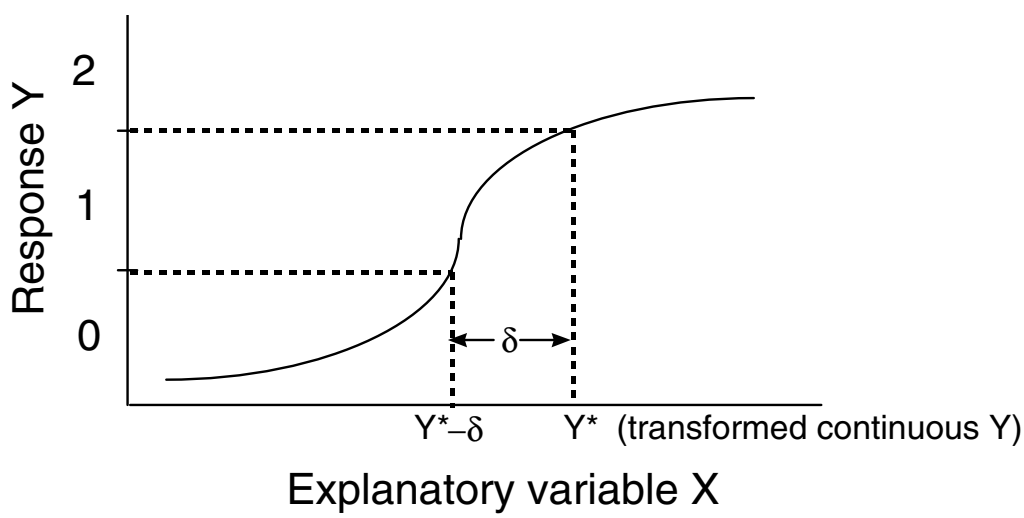


Figure 15.3 Diagram of continuous variable Y* underlying a discrete response variable

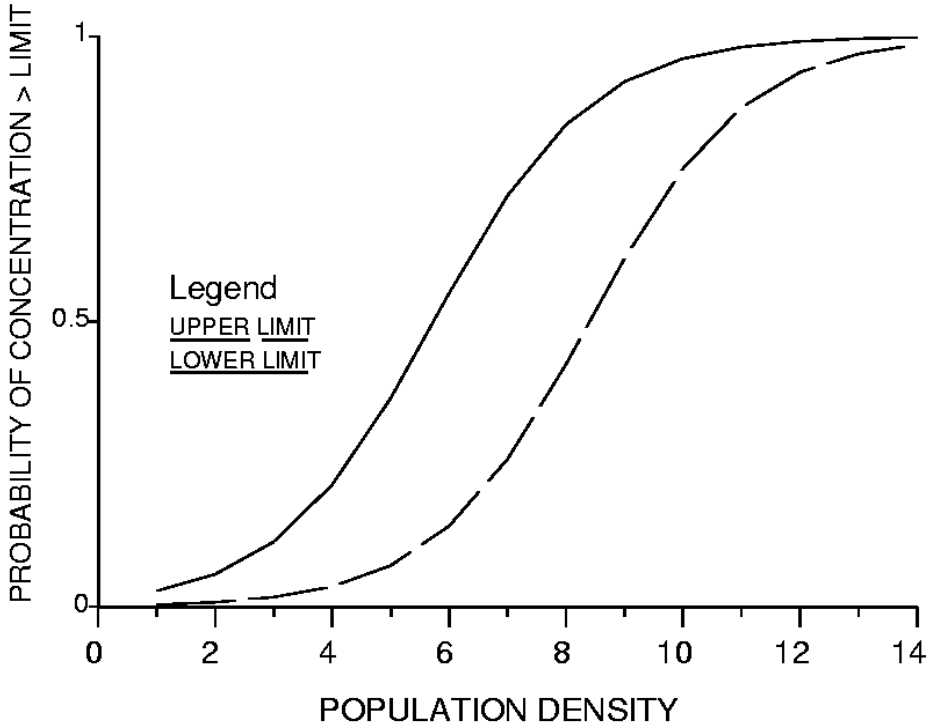


Figure 15.4 Offset logistic curves for an ordered response variable.

15.4.2 Nominal Response Categories

For the situation where there is no natural ordering of the possible response categories, an independent logistic regression must be performed for each possible contrast. Thus if there are m response categories there must be m−1 logistic regressions performed. Coefficients of intercept and slope are estimated independently for each. The econometrics literature has treated this situation in depth -- see for example Maddala (1983). Econometrics categories are often ones of choice -- to purchase one product or another, etc. Examples of unordered variables for water resources applications are not as obvious. However, an understanding of the equations appropriate for nominal responses is important, because these are used when most commercial software is employed to perform logistic regression of ordinal responses.

When independent logistic regressions are computed to determine the likelihood of being below versus above adjacent pairs of categories, no requirement of constant slope across thresholds is made. The probabilities employed may take several forms, but the easiest to interpret are logits of the cumulative probabilities of being below versus above each of the m−1 thresholds

$$\log\left(\frac{\sum \text{Prob } (y > i)}{\sum \text{Prob } (y \le i)}\right), \quad i = 1 \text{ to } m-1. \qquad [15.15]$$

These are called **cumulative logits**, as discussed by Agresti (1984) and Christensen (1990).

For the situation of m=3 ordered responses (0, 1, and 2) corresponding to two thresholds (reporting limit and health standard), m−1 or two logistic regressions must be performed. One equation determines the probability of being at least 1 -- the probability of being above the reporting limit:

$$L_1 = \log\left(\frac{\text{prob}(y=1)+\text{prob}(y=2)}{\text{prob}(y=0)}\right) = b_0 + b_1 X \ . \qquad [15.16]$$

A second equation describes the probability of being at least 2 -- the probability of being above the health standard:

$$L_2 = \log\left(\frac{\text{prob}(y=2)}{\text{prob}(y=0)+\text{prob}(y=1)}\right) = b_0' + b_2 X \ . \qquad [15.17]$$

Together, these two equations completely define the three probabilities as a function of the k explanatory variables X.

Example 1, cont.

Suppose a second threshold at 10 μg/L were important for the TCE data of Eckhardt et al. (1989). This could represent an action limit, above which remedial efforts must be taken to clean up the water before use. Separate logistic regressions were performed for the probabilities of being above the 3 μg/L reporting limit and the 10 μg/L action limit. A new binary response variable, 0 if TCE concentrations were below 10 and 1 if above, was regressed against population density. The results are reported in table 15.6, and the curves plotted in figure 15.5. Note that the two curves are not simply offsets of one another, but have differing slopes. This situation could be viewed as an interaction, where the rate of increase in probability with unit X is not the same for the two thresholds.

Response category	b_0	b_1	lr_0
Above 3 μg/L report. limit	−2.80	0.226	77.0
Above 10 μg/L action limit	−3.37	0.164	23.9

Table 15.6 Independent logistic regressions for two TCE thresholds.

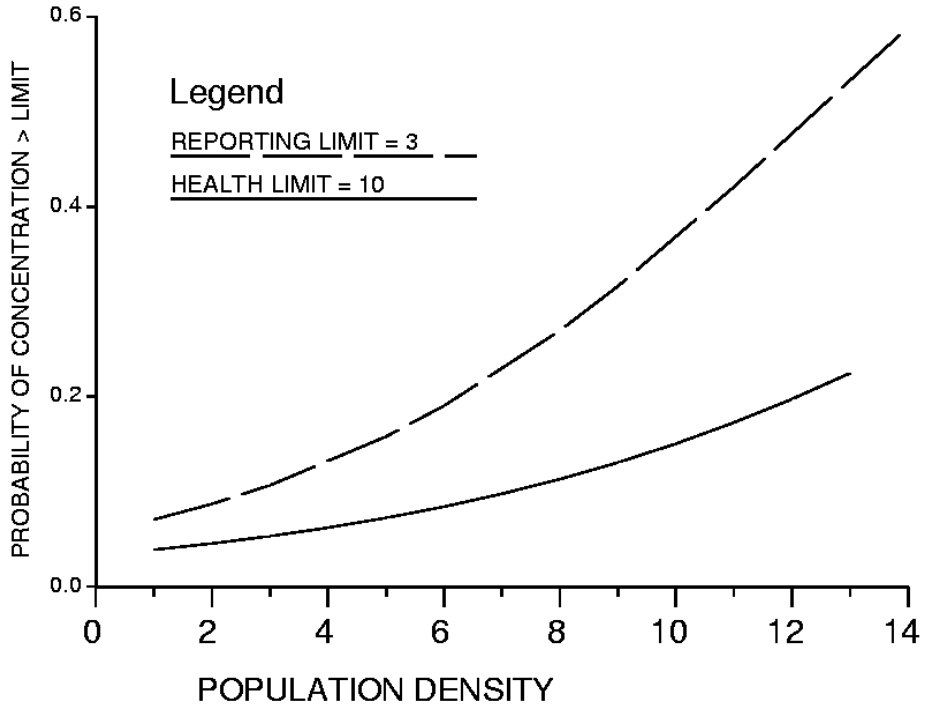


Figure 15.5 Independent logistic curves for two TCE thresholds.

Exercises

15.1 Person and others (1983) evaluated the ability of four factors to predict whether a surface impoundment was contaminated or not. Of particular interest was which of the four factors, information for which must be collected in other areas in the future, showed ability to predict contamination. The factors were:

Factor	Possible scores
Unsaturated Thickness	0 (favorable) to 9 (unfavorable)
Yields: aquifer properties	0 (poor) to 6 (good)
Groundwater Quality	0 (poor) to 5 (excellent)
Hazard Rating for Source	1(low) to 9 (high)

Each impoundment was rated as contaminated or uncontaminated. Using the data in Appendix C20, compute a logistic regression to determine which of the four explanatory variables significantly affects the probability of contamination. What is the best regression equation using one or more of these variables?

Chapter 16

Presentation Graphics

The data are collected, the analyses performed, the conclusions drawn, and now the results must be presented to one or more audiences. Whether by oral presentations or written reports, more information can be quickly conveyed using graphs than by any other method. A good figure is truly worth a thousand table entries.

For oral presentations, rarely are tables effective in presenting information. Listeners are not familiar with the data, and have not poured over them for many hours as has the presenter. Numbers are often not readable further back than the second row. Instead, speakers should take the time to determine the main points to be illustrated, and construct a figure from the data to illustrate those points prior to the presentation. This both shows courtesy to the listeners, and convinces them that the data do provide evidence for the conclusions the speaker has reached.

In a written report, major conclusions are usually listed at the end of the final section, or at the front in an executive summary. A figure illustrating each major conclusion should be contained somewhere in the report. The reader should be able to quickly read an abstract, look at the figures, and have a good idea of what the report is about. Figures should be a "visual abstract" of the report, and are one of the best ways to convince someone to take enough time to read your work. They again give evidence that the data do support the conclusions you have reached.

All graphs are not created equal. Some present quantitative information clearly and precisely. Others are not as effective, and may even be misleading. Guidelines to the "level of precision" for common types of graphics are presented in this chapter. Also presented are a collection of misleading graphics which should be avoided. These come largely from experience, driven by the impression that their use is becoming more common in graphics software on microcomputers.

Understanding the strengths and weaknesses of various types of graphs is important when choosing the most appropriate way to present data. Three references stand out in their evaluation of graphs for quantitative data: Cleveland (1985) discusses the ability of the human eye-brain system to process information. Tufte (1983) describes the artistry involved in creating graphics. Schmid (1983) is a handbook listing numerous examples of both good and bad graphics. This chapter draws on ideas from these three and others.

16.1 The Value of Presentation Graphics

Graphs can clarify complex interrelationships between variables. They can picture the "signal" over and above the "noise", letting the data tell its story. In Chapter 2, graphs for understanding data were discussed. These same methods which provide insight to an investigator will also illustrate important patterns and contrasts to an investigator's audience.

Tables simply do not allow easy extraction of a data signal. For example, Exner and Spalding (1976) and Exner (1985) determined concentrations of nitrate in about 400 wells in the Central Platte region of Nebraska ten years apart -- in 1974 and 1984. As little information is available about changes in groundwater quality over time, these are important studies. Data were displayed with maps and tables for each separate period. Comparisons between the periods were done as narrative text, relying on the tables and maps. To better illustrate these data, lowess smooths of nitrate concentration versus depth for the two time periods are shown in Figure 16.1. This concise figure effectively illustrates the increases in nitrate at a given depth over the ten year period, and the decrease in concentration with depth. It shows that increases in concentration over the 10 years are much larger at shallow depths. For a specific nitrate "action level" such as 8 mg/L, the increase in depth reached on average by this concentration can be calculated. Perhaps the valley of lower concentration for the shallow system evident in both time periods can be explained by physical factors, leading to an important new understanding. Or perhaps the wells sampled at these depths were different in some characteristic, leading the scientist to sample additional wells more like those at other depths. A good graph will provide much more understanding than a table to the audience, whether they are scientists or managers, often leading to new understanding or to better decisions.

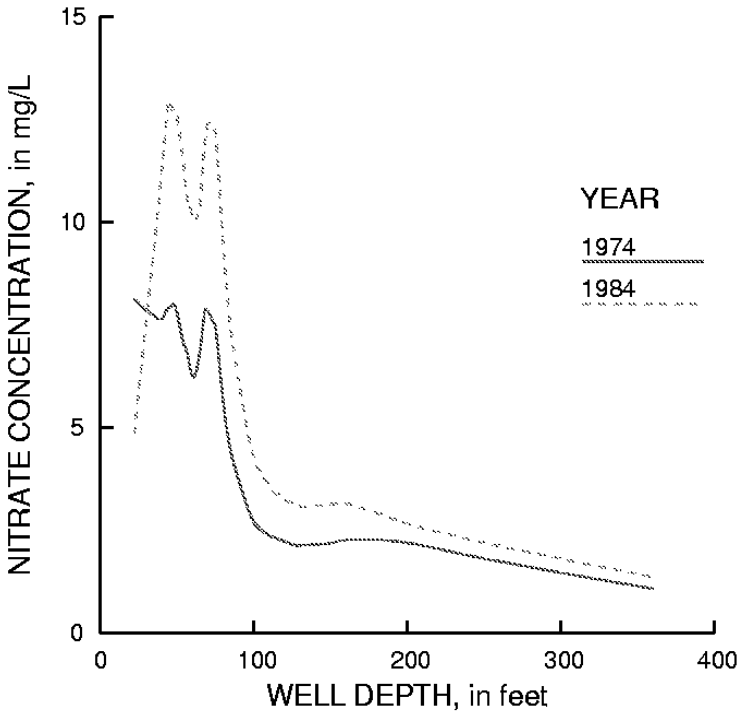


Figure 16.1 Nitrate concentrations in Nebraska groundwater.
Data from Exner and Spalding (1976) and Exner (1985).

16.2 Precision of Graphs

The purpose of a scientific graph is to display quantitative information in a clear and concise manner, illustrating a major concept or finding. During the 1980s research was conducted to determine how easily the human eye-brain system can perform various tasks of perceiving and processing graphical information. The purpose was to rank tasks necessary in interpreting common graphs, such as bar and pie charts, in order to understand which types of graphs are most effective in presenting information. Prior to this time scientists had no objective means of determining which graphs should be preferred over others, and choice was merely a matter of preference.

The primary study was conducted by Cleveland and McGill (1984a). Their major precept was stated as:

> A graphical form that involves elementary perceptual tasks that lead to more accurate judgments than another graphical form (with the same quantitative information) will result in better organization and increase the chances of a correct perception of patterns and behavior (pages 535-6).

They then ranked perceptual tasks on the basis of accuracy, as determined by the number of correct judgments of identical data displayed by different graphs. This ranking is given in Table 16.1. Their concept of accuracy might also be thought of as precision -- smaller trends or differences between data groups can be discerned using more "accurate" tasks. Use of graphs employing tasks higher in table 16.1 will allow smaller differences or trends to be seen. Tasks lower in the table are sufficient to display only larger differences. These lower tasks are those most commonly found in "business graphics", newspapers, and other popular illustrations. Thus when deciding which types of graphs to use, both the precision needed and the expected audience must be considered. When scientists are the main audience, graphs using tasks as high in the table as possible are preferable. When less precision is required to illustrate the main points and the audience is the general public or managers, some of the less precise business graphics may communicate more easily.

More Precise	Position along a common scale
•	Positions along nonaligned scales
•	Length, Slope, Angle
•	Area
•	Volume, Curvature
Less Precise	Shading, Color saturation

Table 16.1 Precision of perceptual tasks (adapted from Cleveland and McGill, 1984a).

16.2.1 Color

Color can both enhance and interfere in the ability to precisely and accurately read graphs. It can interfere in judgments of size between areas of different colors (Cleveland and McGill, 1983). From color theory it is known that "hotter" colors such as reds and oranges, and colors of greater saturation will appear larger than "cooler" colors (blues) and pastels (lesser saturation). Therefore areas shaded a bright red on a map, as is commonly done for computer-map output of pollution studies, will appear larger than they would if shaded another color or with a pastel such as light pink. The eye is drawn to these areas, and their impression is larger than the proportion they would receive by area alone.

Pastels can therefore be used to minimize the biasing effect of both hotter and brighter colors. The low saturation ("washed-out" color) minimizes differences between hotter and cooler shades, and therefore put all areas on an equal footing. Of course this defeats the "newspaper graphics" effect of attracting attention to the graph, but enhances the graph's ability to portray information.

Color can also be quite helpful in presenting data when judgments of size are not being made. When differentiating groups of data on a graph, for example, each group could be assigned a

different color, as opposed to a different symbol or letter. Circles or dots of differing colors allow greater visual discrimination than do differing symbols or letters (Lewandowsky and Spence, 1989). Similarly, color lines allow better perception than solid versus patterned lines. As color is not yet widely available in scientific publication media, its best use to date is in presentations at conferences and lectures. Here color can greatly aid the viewers' precision in differentiating points and lines representing data of different groups.

16.2.2 Shading

Figure 16.2 illustrates the most common use of shading -- shaded maps where the density of the ink indicates the magnitude of a single variable. The maps may be of the entire country, a state, or a study area. These "shaded patch maps" or "statistical maps" have inherent difficulties for correct interpretation.

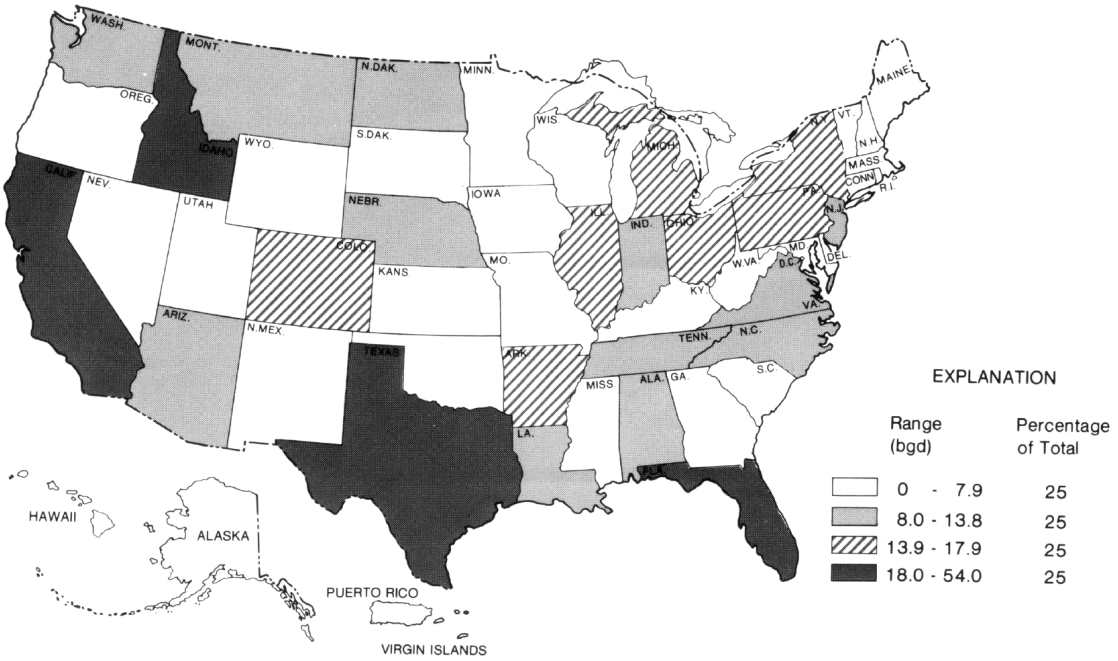


Figure 16.2 Total offstream water withdrawals by state, from Solley et al. (1983).

The first difficulty is that the impression an area makes on the human brain is a function of both the shading and the size of the polygon. Thus larger areas stand out in comparison to smaller areas, though their shading may be equal. In figure 16.2, Texas stands out not only because it is dark, but because it is large. Of the lightly shaded states, the eye is drawn to Montana (MONT) because of its size rather than to New Jersey (NJ). However, an area's importance may not be related to its physical size. If population is important, as it may be for the water withdrawals in each state as shown in figure 16.2, a state with a higher population like New Jersey may be far more important than is Montana, a state with much smaller population. The weighting given to larger areas on a shaded map is often inappropriate to the data being illustrated.

A second limitation is that all variability within areas is totally obscured. Thus a map is only as precise as the size of the areas being shaded. Water use undoubtedly varies dramatically across Texas and other states, but that cannot be shown on a shaded map unless the states are subdivided into counties. Counties vary considerably in size across the country, so that the generally larger counties in the Western U. S. will produce greater impressions on the viewer than do smaller Eastern counties.

Third, only a small number of shading levels can be distinguished on a map. Five shades of grey including black and white can usually be portrayed, but more than five is difficult to distinguish. Differences degrade as graphs are reproduced on a copier. In an attempt to augment the number of classes shown on a map, patterns of lines and cross-hatching are sometimes used, such as the 13.9-16.9 class in figure 16.2. Such patterns quickly become very confusing, actually reducing the eye's ability to distinguish classes of data. One must also be careful to use a series of patterns whose ink density increases along with the data Figure 16.2 seems to violate this rule, as the shade of the second class (8.0-13.8) appears darker than the third striped pattern.

Two types of alternatives to shaded maps are tried. The first type continues to display the geographic distribution on a map, with symbols depicting data classes within each area (each state). Circles or squares with shading or color according to the classification are one possibility. Bars are another possibility (figure 16.3). With bars the perceptual task is a judgment of length without a common datum, an increase in precision in that differences between areas may be distinguished at more than five levels. However it is often difficult to place the bars within state boundaries. Framed rectangles (figure 16.4) are another symbol which may be used within each state. For these the task is a judgment of length along a non-aligned common scale, an improvement in precision over judgments between shadings.

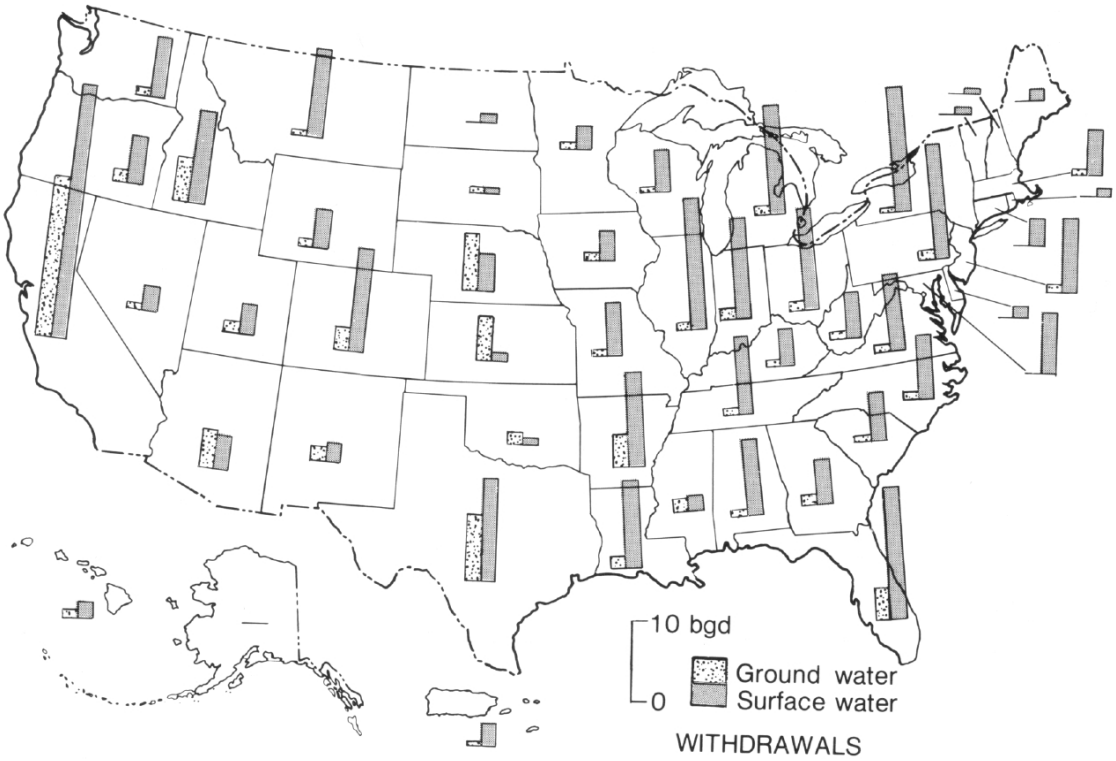


Figure 16.3 Withdrawals for offstream use by source and state, from Solley et al. (1983).

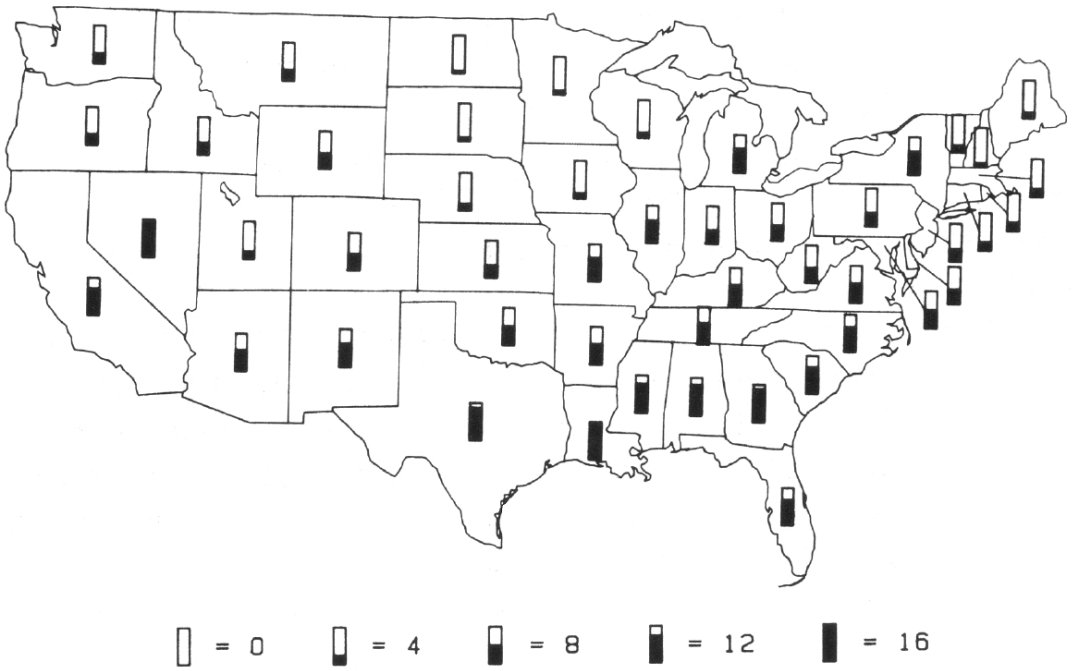


Figure 16.4 Murder rates per 100,000 population, from Cleveland and McGill (1984a).

© American Statistical Association. Used with permission.

The second alternative to shaded maps is to abandon a map background, and construct bars or other ratings for each state. These can be classed by region, though much of the regional perspective is sacrificed for state-by-state precision when abandoning maps.

16.2.3 Volume and Area

The most common use of area perception is with pie charts. These graphics are most often used when the sum of data equals 100 percent, so that slices of the pie indicate the relative proportion of data in each class (figure 16.5). However, only large differences can be distinguished with pie charts because it is difficult for the human eye to discern differences in area. In figure 16.5 it is only possible to see that the northeast slice in the lower right part of the pie is larger than the others. No other differences are easy to distinguish.

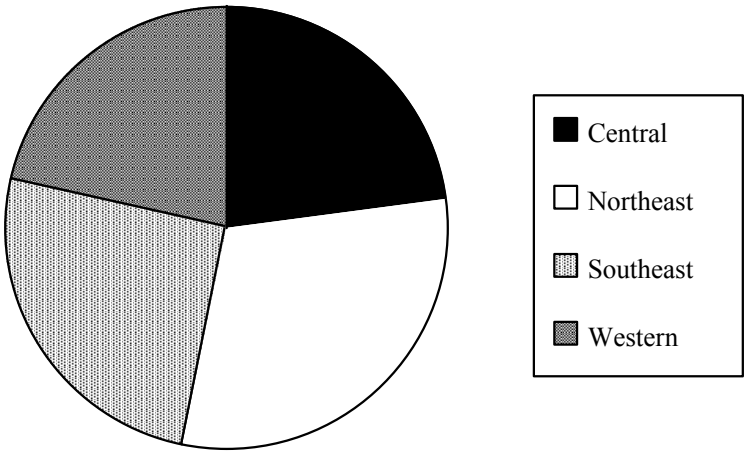


Figure 16.5 Numbers of students from four regions of the U. S.

It is always possible to replace a pie chart with a figure using one of the higher perceptive tasks in order to improve precision. For the same data of figure 16.5, figure 16.6 presents a "dot chart" (Cleveland, 1984), a thin bar graph. Now the judgment is of location along a common scale (the y-axis), and all differences are clearly seen. The four regions can be ordered and estimates of the magnitude for each read from the scale. The data are displayed with much greater precision than with a pie chart.

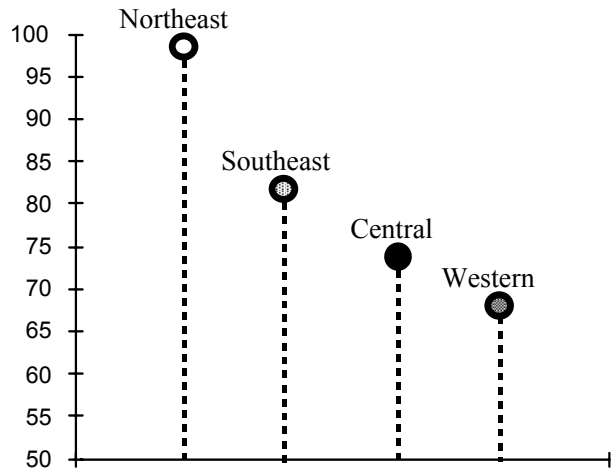


Figure 16.6 Dot chart of the student data of figure 16.5.

Pie charts have little utility for scientific publications, due to their imprecision. The comparison of water quality at two stations in figure 16.7, for example, would be better done using a more precise method, such as two stiff diagrams (see Chapter 2). The presence of numbers on the graph is a clue that the graph is incapable of portraying differences with the necessary precision. It is instead a circular table. Graphs with numbers are often a "red flag", signalling the inadequacy of the graph itself.

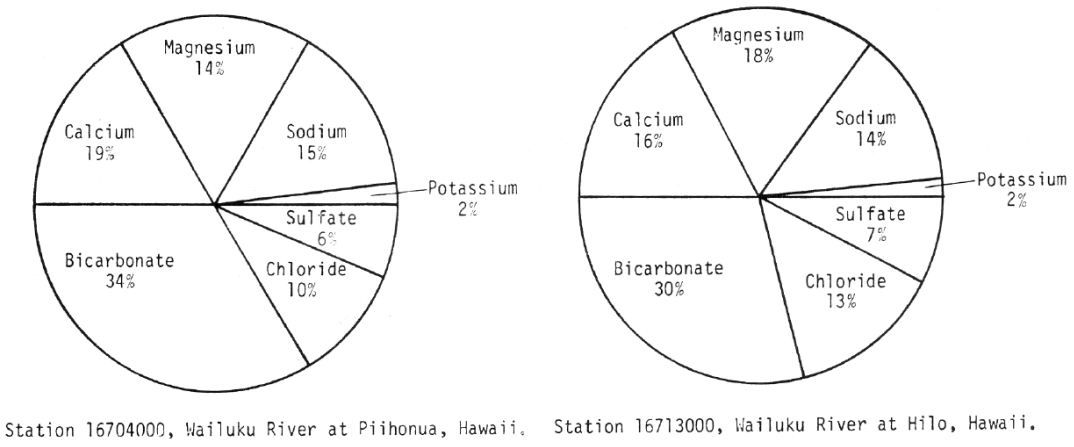


Figure 16.7 Water quality at two sites in Hawaii (from Yee and Ewart, 1986).

16.2.4 Angle and Slope

Judgements of angle and slope occur when comparing two curves, such as in figure 16.8. Differences between the curves are often of interest, and differences are represented as distances in the y direction. However, the human eye sees differences primarily in a direction perpendicular to the slope of a curve, much like the least normal squares line of Chapter 10. We

do not naturally see differences as they are plotted. So in figure 16.8 it appears that differences are largest in the center, and smallest at the extremes of X. However, the bottom figure shows the differences directly. The largest differences are on the left, with a linear decrease as X increases! To truly see differences in the top figure a judgment is required about the angles of the lines in relation to the y axis, and this is quite difficult. A good rule of thumb is that if differences are of interest, plot the differences directly.

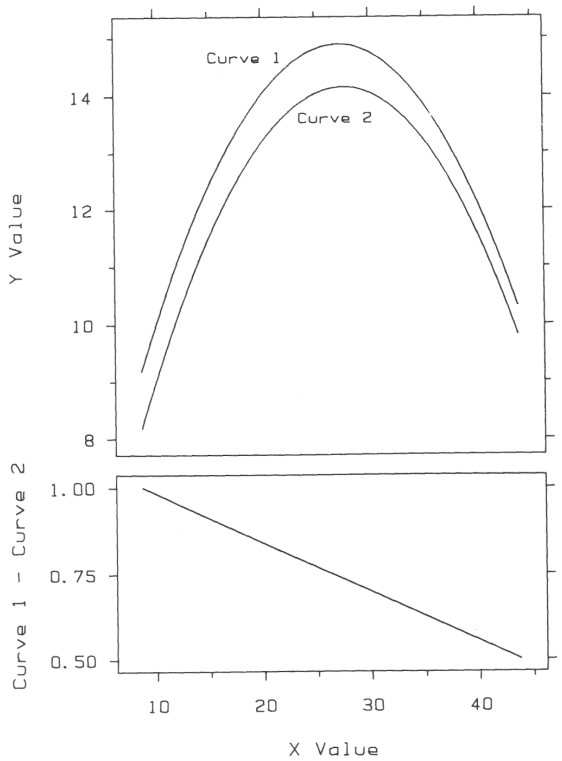


Figure 16.8 Comparison of two curves. From Cleveland and McGill (1985).

© AAAS. Used with permission.

Figure 16.9 is a comparison of measured and model logarithms of streamflow. Which days show the poorest predictions? Though it appears that the largest difference in log streamflow occurs on May 16 and in latter June, the mismatch is actually much greater on and near May 6. If the purpose of the graph is to portray daily differences, the differences themselves should be plotted.

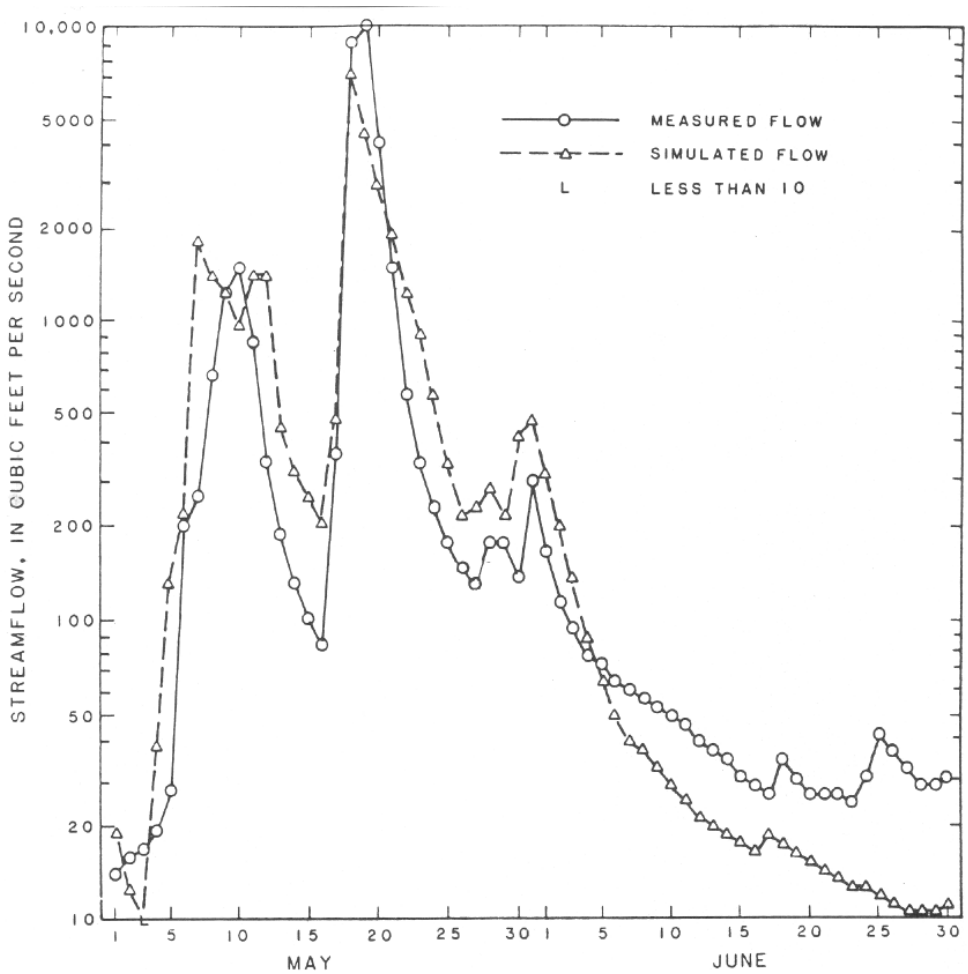


Figure 16.9 Measured and simulated streamflow. From Bloyd et al., 1986.

Another type of graph which uses judgments of slope and angle is a cumulative line graph such as figure 16.10. A quick look at the graph might indicate that x2 and x3 are increasing, simply because their baseline is increasing. To determine the magnitude of any variable except the one whose base is the x-axis requires compensating for the non-horizontal baseline angle as it changes across the range of X. This is obviously difficult to do. The determination of which of the three items in figure 16.10 is largest in periods 1 and 2 is also quite difficult, for example.

One justification for cumulative line graphs is that they show the proportion of values against the total, which is shown as the top line. Moving up the table of perceptual tasks results in a better solution -- to plot each of the variables separately, and plot the total if it is important. This is done in figure 16.11. Determination that x3 is either equal or greater than the others during periods 1 and 2 is much easier here. The cyclic variation of x3 is also easier to spot. Comparisons between variables with small magnitudes such as x2 and x3 are not swamped out by larger variations in the variable at the base (x1). Judgements are made using position along a common scale (the y-axis), a much easier and more precise task than in 16.10.

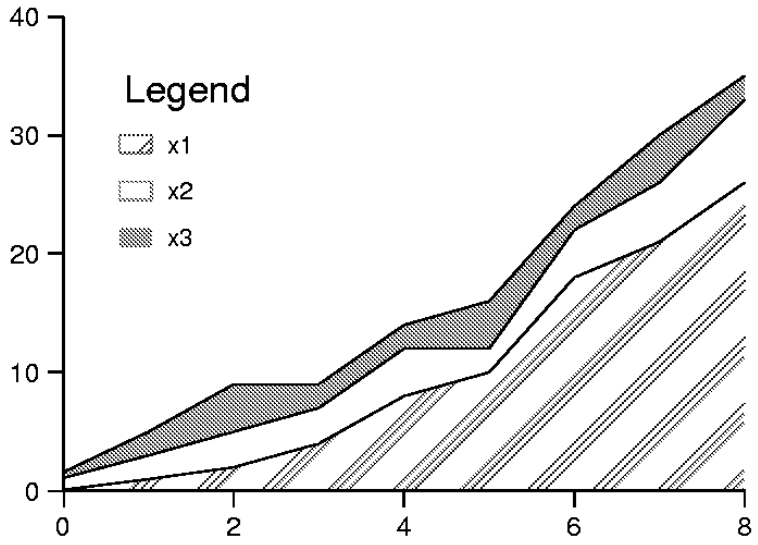


Figure 16.10 Cumulative line graph.

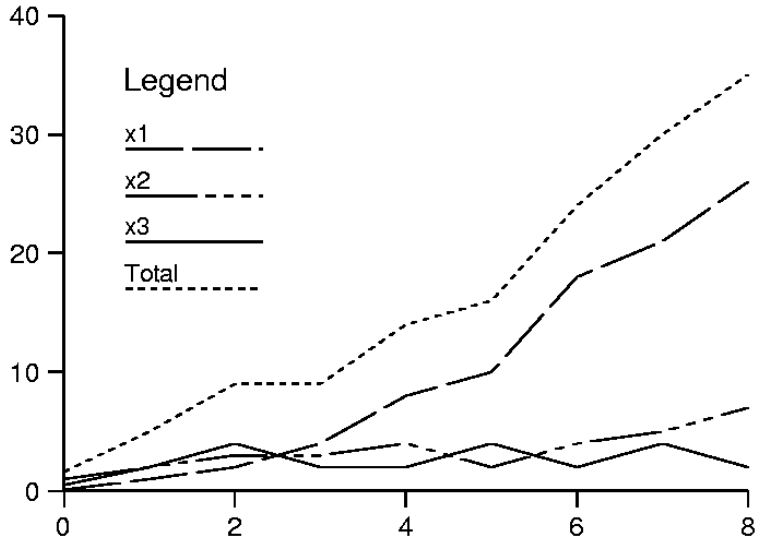


Figure 16.11 Variables of figure 16.10 plotted individually.

16.2.5 Length

Judgements of length are required when symbols or bars are to be measured which do not have a common datum, and where no common scale is available. Figure 16.12 shows the simplest such case, determination of the length of two offset bars. Which is longer is difficult to visually

determine. An example requiring the use of length judgments are the bars displayed on the map of figure 16.3.

To make these judgments more precise, a common scale can be added to each bar. This is done in figure 16.13 as a framed rectangle. The rectangle surrounding each bar is of exactly the same length, a common reference frame. It is now easier to see that the first bar is indeed longer than the second. This is because the judgment is made using positions of the white areas within the common scale. Their relative differences are greater than the shaded bars, and so more easily seen. In situations where a common datum is impossible such as multiple stiff or other diagrams located on a map, adding a frame of reference will improve the viewer's precision in discerning differences.

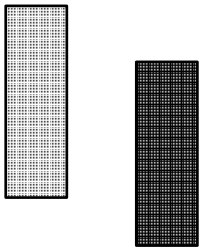


Figure 16.12 Judgement of length without a common scale or datum.

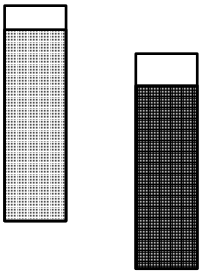


Figure 16.13 Framed rectangles of the figure 16.12 data, adding a common scale.

16.2 6 Position Along Nonaligned Scales

Framed rectangles are examples of graphs with a common but nonaligned scale, ie. without a common datum. Another graph in this category is a stacked bar chart: stacked (figure 16.14). These graphs of segmented bars require judgments of position and/or length. Only the lowest segments of each bar possess a common datum --they are the easiest to compare. All other comparisons between bars, and among segments within a bar, are more difficult without a common datum. For example, in figure 16.12 it is difficult to determine which of the top two

squares of bar 1 is larger. How about the top and bottom squares (D vs A) of bar 3? Or group B squares for bars 1 and 3?

To make comparisons more precise, stacked bars can always be unstacked and placed side-by-side, producing grouped bar charts (figure 16.15). These graphs belong in the highest precision category -- position along a common scale (common datum). By using a common datum, smaller differences are more easily seen. For example. in bar 1 it is now easy to see that C is larger than D. Square A is larger than D in bar 3, and the group B square for bar 1 is larger than bar 3. The precision with which the graph can be read is greater for the grouped bar chart than the stacked chart, a distinct advantage.

Often bars are stacked so that their totals are easy to compare. With grouped bar charts this is easily accomplished by plotting separate bars of group totals. As both types of bar charts are equally familiar to viewers, it is difficult to see why stacked bars should ever be used over grouped bar charts.

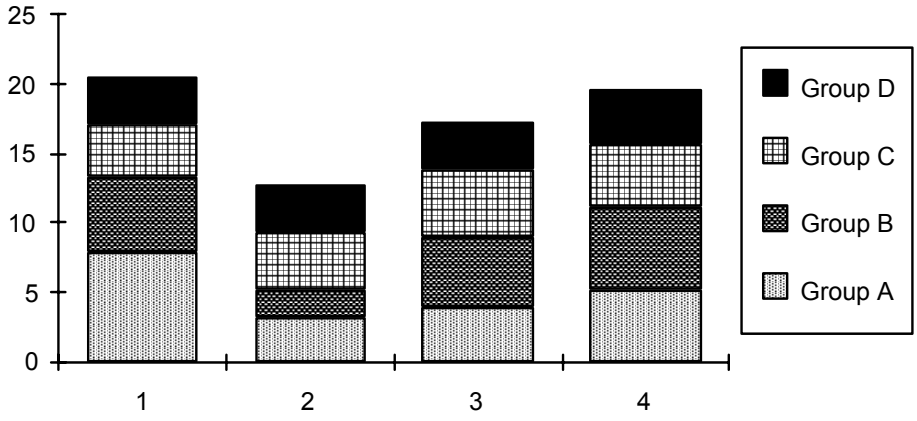


Figure 16.14 Stacked bar charts.

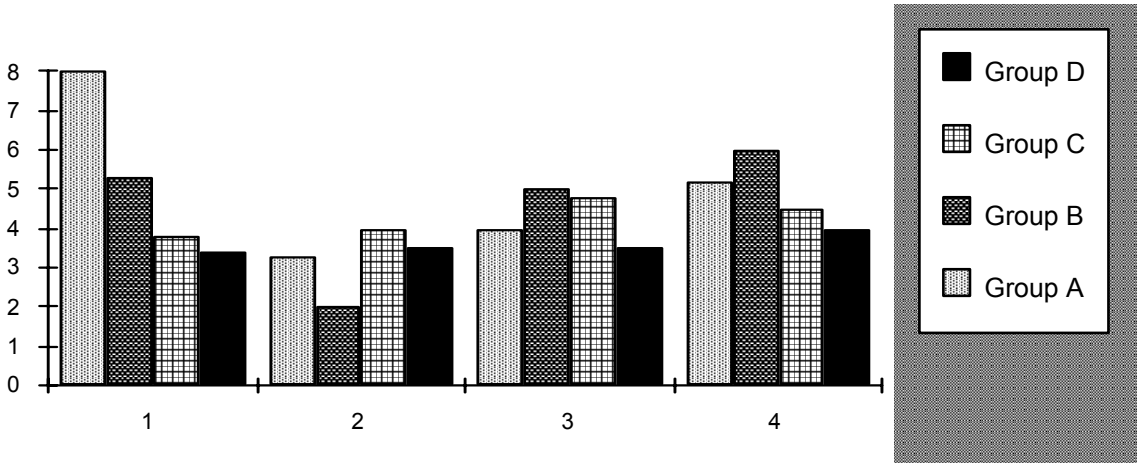


Figure 16.15 Grouped bar charts for the figure 16.14 data.

16.2.7 Position Along an Aligned Scale

Grouped bar charts are one example of graphs where data are shown as a position along an aligned (common datum) scale. Also in this category of highest precision are the dot charts of Cleveland (1984). These "skinny bar charts" (figure 16.16) remove some of the visual confusion of bar charts due to the area and shading of the bars. The dots highlight the only information present -- the position at the top of the bar. Though for simple situations the two graphs are equivalent, for complex situations dot charts more clearly show the information. A final advantage to dot charts are that error bars around each value can easily be added.

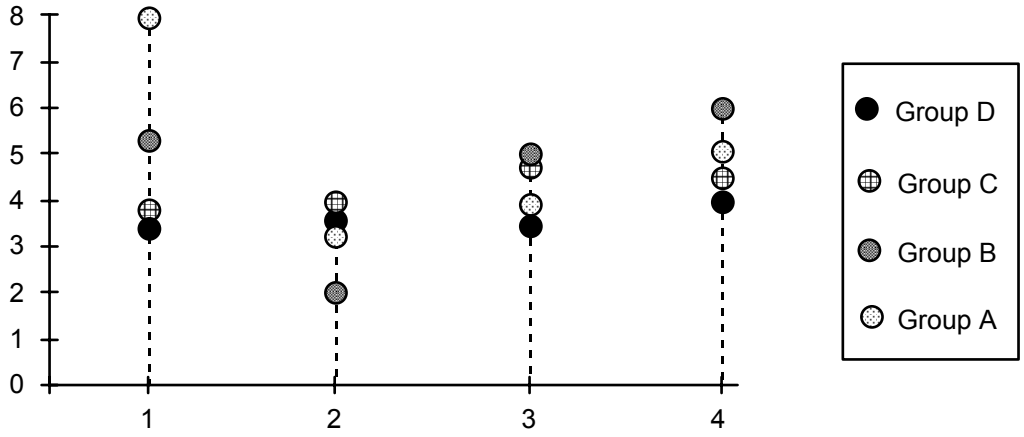


Figure 16.16 Dot chart for the figure 16.15 data.

Other more-familiar types of graphs also belong in this category, including scatterplots and boxplots. Though discussed and used fully throughout this book, the strengths of boxplots bear repeating. Many boxes can be placed on a page, allowing precise summaries and comparisons of a large amount of information. In figure 16.17 are boxplots for a two-way ANOVA situation. Differences in concentration due to both land-use category and to sewering are easily seen, as are skewness and outliers, by comparing the boxes.

16.3 Misleading Graphics to be Avoided

16.3.1 Perspective

Figures are often put into perspective, that is tilted to give an impression of three dimensions, in newspaper and other popular graphics. The intent is to make the figure look more "solid". Unfortunately by doing so, judgments of area, length and angle used by the viewer to extract information become impaired. Numerical values are altered by the tilting so that they no longer can be accurately read. Thus the appearance may be more solid, but the information is less so.

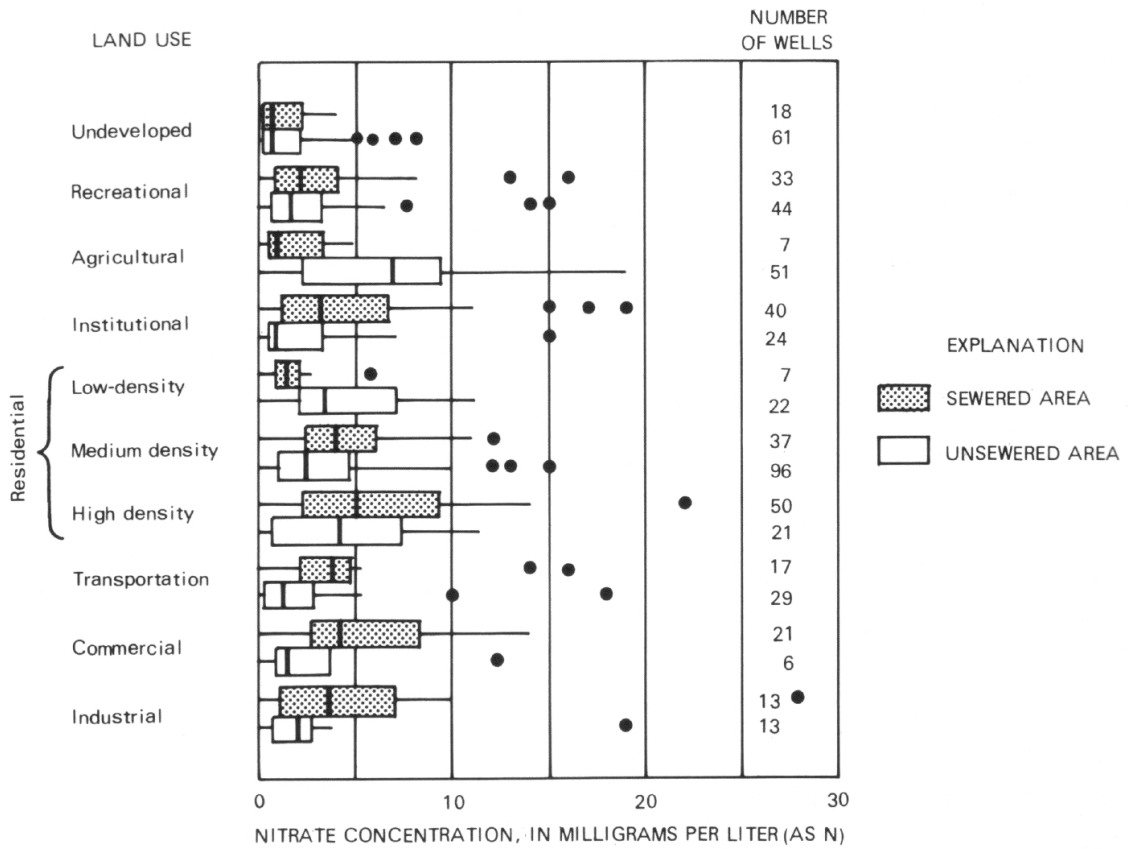


Figure 16.17 Boxplots of nitrate concentrations by land use and sewering. From Eckhardt et al., 1989.

One example is given in figure 16.18. A pie chart is presented on the left with three slices (labeled A, B and C) of exactly equal size. After being put into a perspective view at the right, the slices no longer appear equal. Judgements of angle such as these are impossible to get correct once the angles are altered by perspective. A second example is figure 16.19. There bar charts are placed into perspective so that lots of bars can be crammed into one figure. A resulting problem is that some bars are hidden by others. A more serious problem is that comparisons of bar heights must be done along a sloping plane. The base of the graph is not level, but increases towards the back. This makes judgments between bar heights difficult. For example, which is higher, the thermoelectric withdrawals for 1965 or irrigation withdrawals for 1970?

Viewers will tend to see bars towards the back as higher than they should in comparison to bars nearer the front when perspective is used to tilt the base. Thus the front of the "thermoelectric" bars must be compared to the back of the "irrigation" bars in order to accurately assess the data portrayed by bar heights. Comparisons of heights across non-adjacent rows is even more

difficult. The two bars cited above have exactly the same value of 130,000 million gallons per day, though the one at the back appears higher.

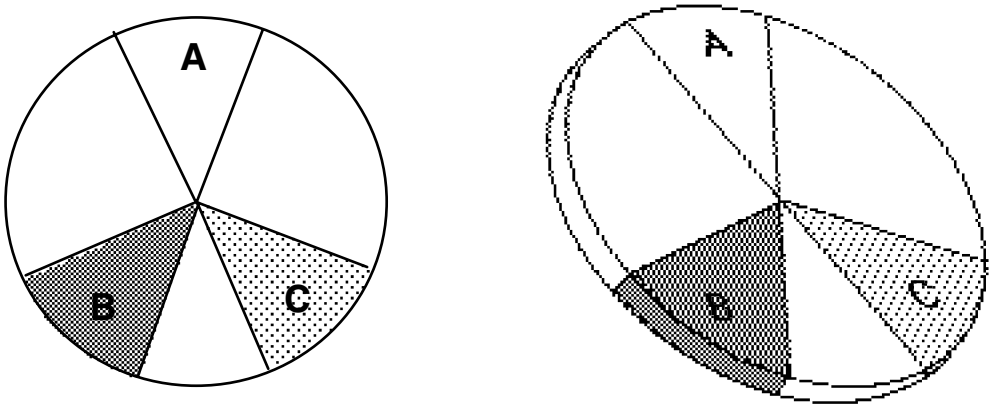


Figure 16.18 Pie chart before and after being placed into perspective view.

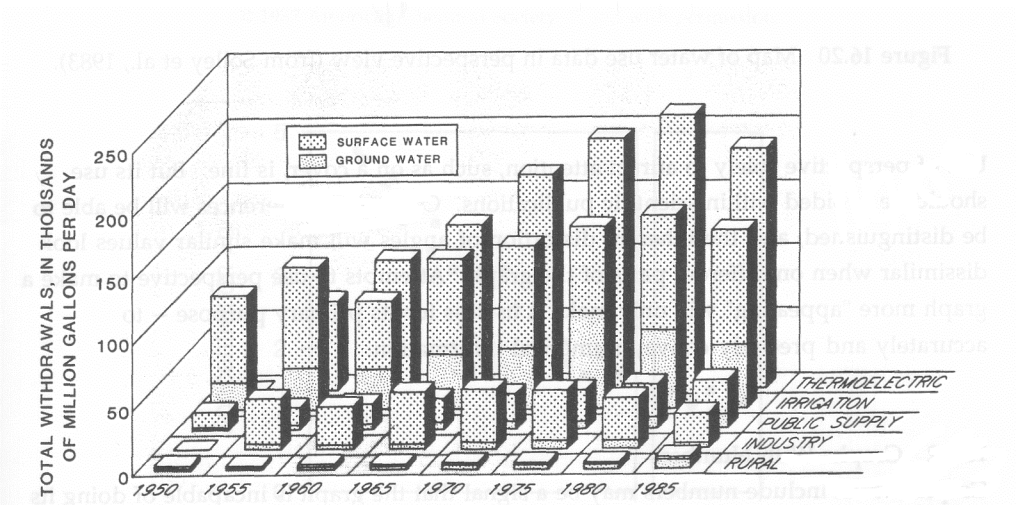


Figure 16.19 Bar chart of water use data, in perspective view (from Solley et al., 1988).

Perspective should also be avoided when presenting maps. Figure 16.20 is a perspective map of water use in the United States (Solley et al., 1988). Because the base of the map is tilted, values at the back will look higher than those in the front for the same quantity. Comparisons between Montana (at the back) and Louisiana (at the front), for example, are quite difficult. From a table inside the report, Louisiana has a larger value, but it doesn't appear that way on the map. Note also that several states are again partially or totally hidden.

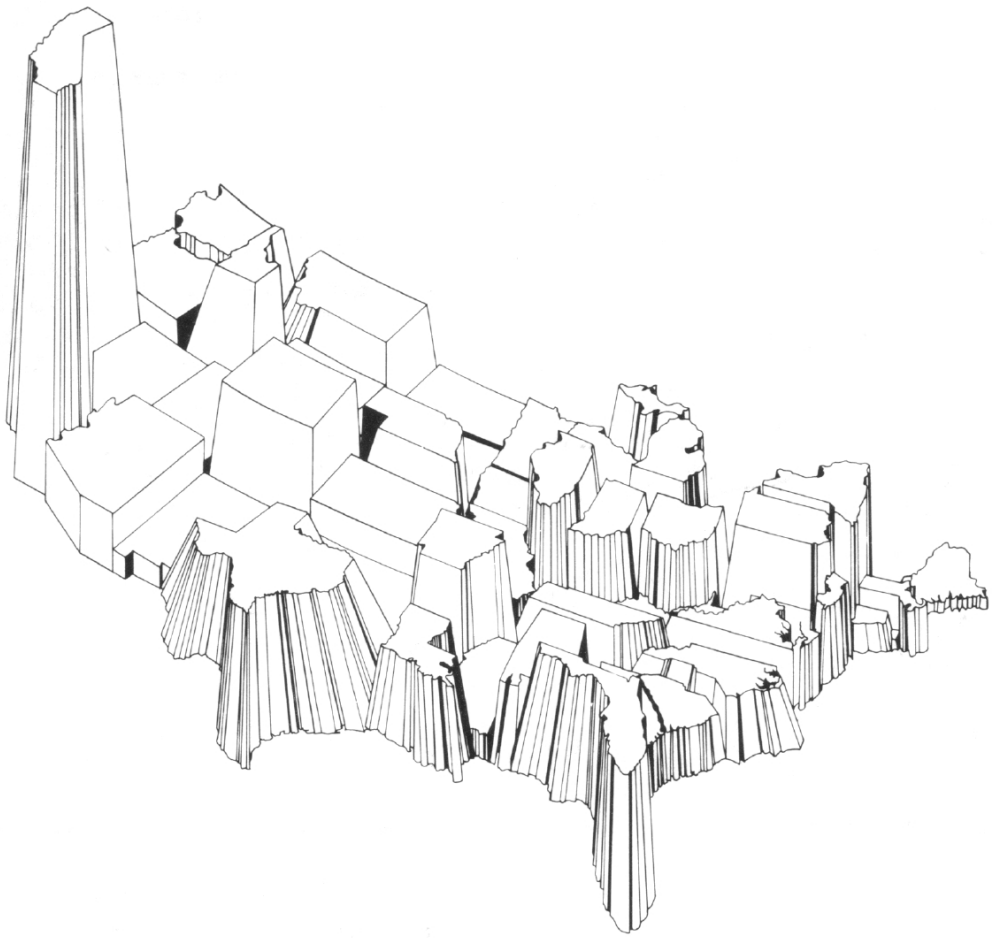


Figure 16.20 Map of water use data in perspective view (from Solley et al., 1983).

Use of perspective solely to attract attention, such as on a cover, is fine. But its use should be avoided within scientific publications. Only large differences will be able to be distinguished, and the inherent distortion of angles will make similar values look dissimilar when on different parts of the graph. Attempts to use perspective to make a graph more "appealing" will only make it useless for its primary purpose -- to accurately and precisely convey numerical information.

16.3.2 Graphs With Numbers

Graphs which include numbers may be a signal that the graph is incapable of doing its job. The graph needs to be made more precise. See for example figures 16.7 and 16.26. Tables providing the necessary detail for computations can be placed elsewhere in the report if required. But they do not provide the insight needed to quickly comprehend primary patterns of the data. Adding numbers to graphs which also do not portray those patterns does not add up to an effective graph.

16.3.3 Hidden Scale Breaks

Breaks in the scale of measurement on a graph can be very misleading to the viewer. If scale breaks are used, it is the job of the presenter to make them as clear as possible. For example, the scale break in figure 16.21 is not very obvious (it is also not necessary). Bars are drawn right through the data, incorrectly implying that no break is present.

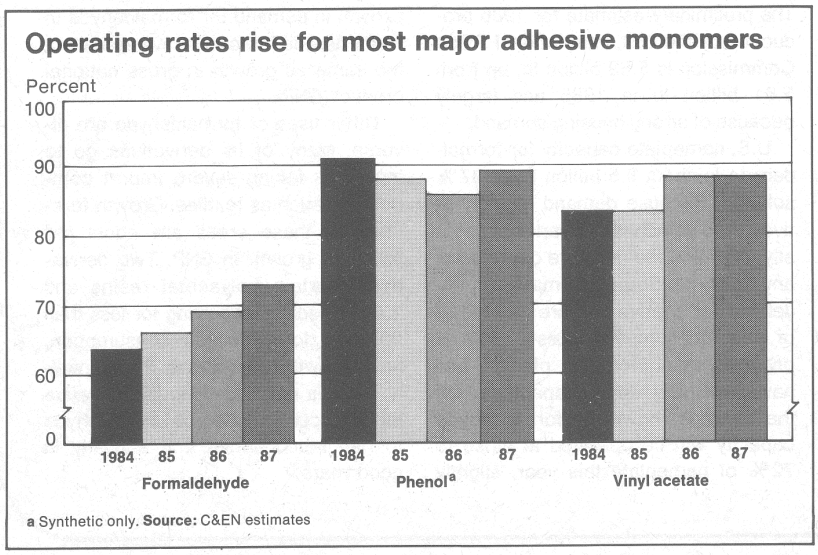


Figure 16.21 Hidden scale break, from Greek (1987)

© 1987 American Chemical Society. Used with permission.

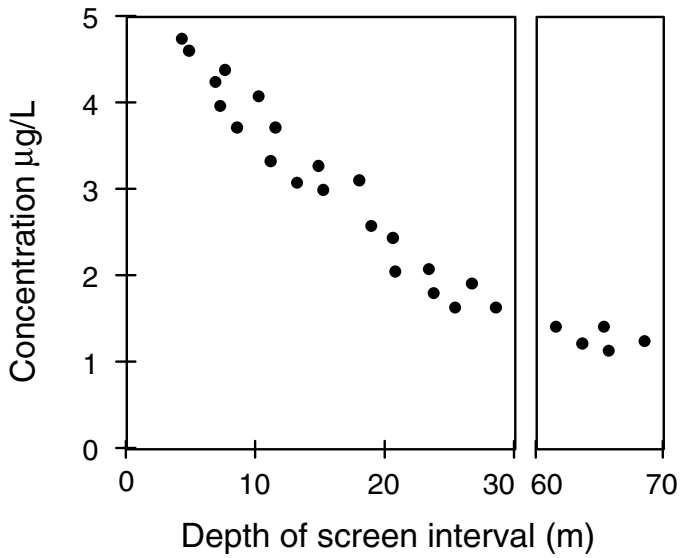


Figure 16.22 Full scale break.

To make a scale break more obvious, Cleveland (1984) suggested the use of a "full scale break" as in figure 16.22. There the jump in depth of wells used for sampling is clearly portrayed. It is difficult for the viewer to misinterpret a scale break using this method. We heartily recommend use of full scale breaks when breaks must be used. Better yet, avoid using scale breaks by employing a transformation of the data such as logarithms to make the break unnecessary.

16.3.4 Overlapping Histograms

Overlapping histograms are one of the worst graphs for comparing groups of data, and yet are quite common. They totally obscure differences and similarities between groups. With the excellent alternative of group boxplots available there is little reason to use them. Figure 16.23 shows two sets of overlapping histograms, effective porosity (A) and infiltration capacity (B). Three groups are being compared in A, and two groups in B. There is no way a reader could verify or disprove any conclusions reached in the report concerning these variables by looking at these histograms. The use of lines for shading, automatically produced by many graphics software programs, only makes matters worse. In general, avoid using overlapping histograms!

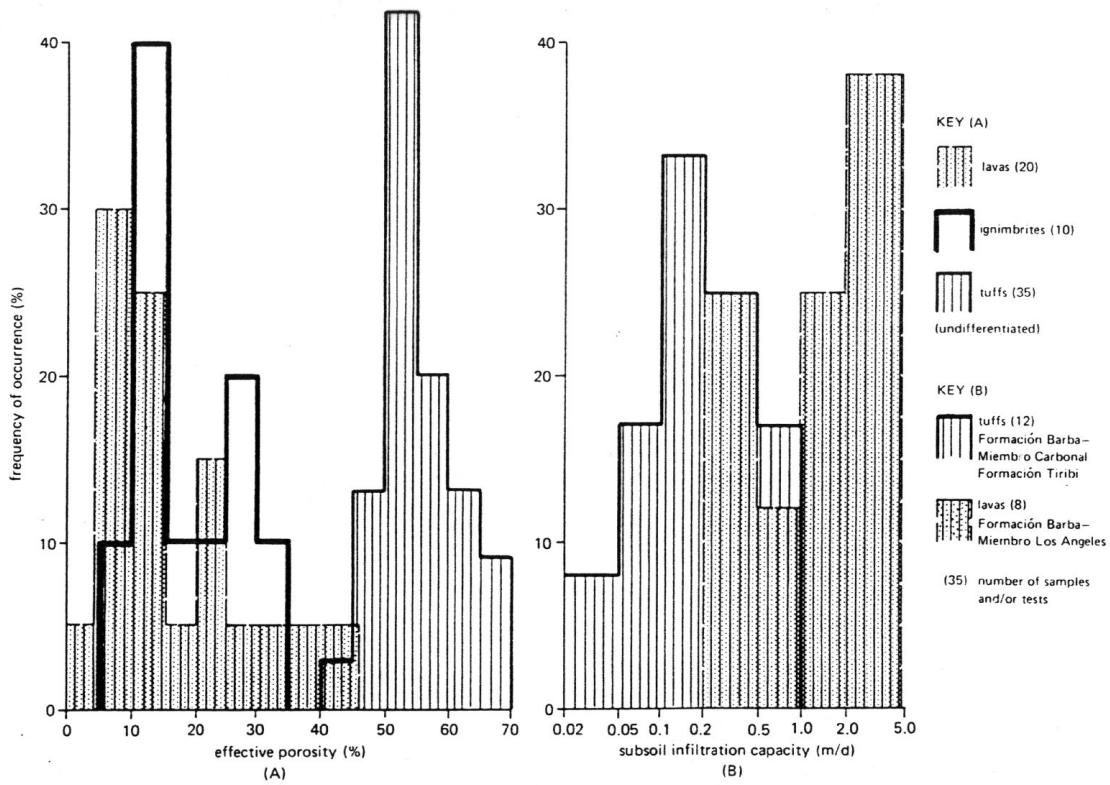


Figure 16.23 Overlapping histograms of two variables. From Foster et al. (1985).

Exercises

16.1 Field and laboratory pH were measured on the same samples by Bachman (1984) to determine if values changed over the time it took for shipment to the lab . The data were plotted in the figure below. How might the graph be improved in order to show this comparison?

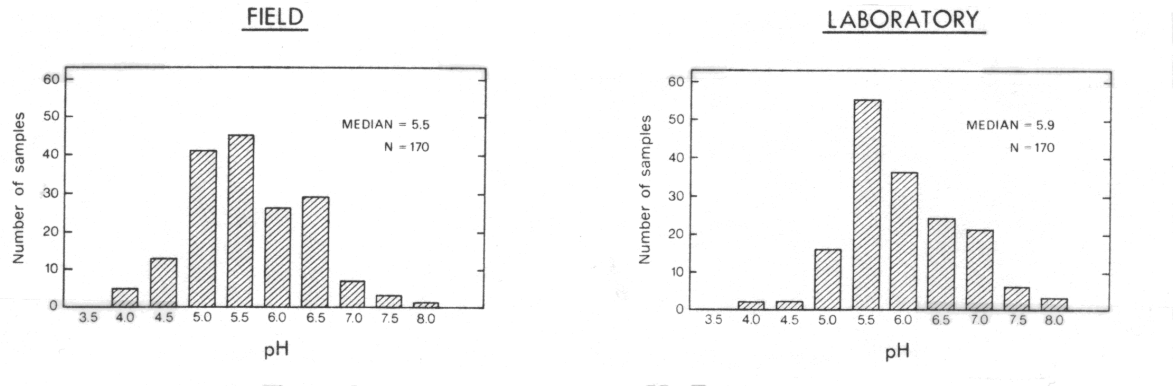


Figure 16.24 Field versus lab pH. From Bachman, 1984.

16.2 Seasonal patterns of specific conductance for stations along the Merced River are shown below. How might this graph be improved to better show both seasonal and downstream differences?

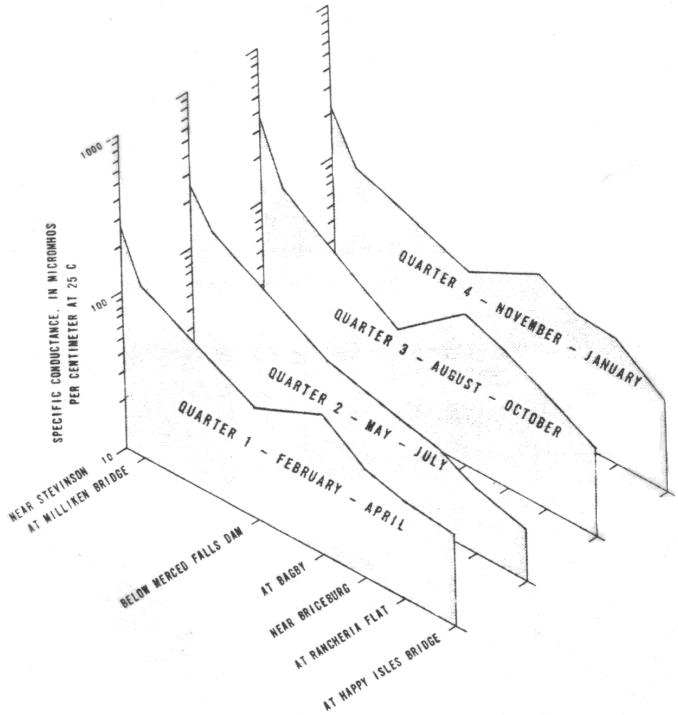


Figure 16.25 Specific conductance along the Merced River. From Sorenson, 1982.

16.3 Variations in dissolved oxygen and biochemical oxygen demand (BOD) were documented along the Trinity River watershed. How might the graph be improved in order to better show differences between sites?

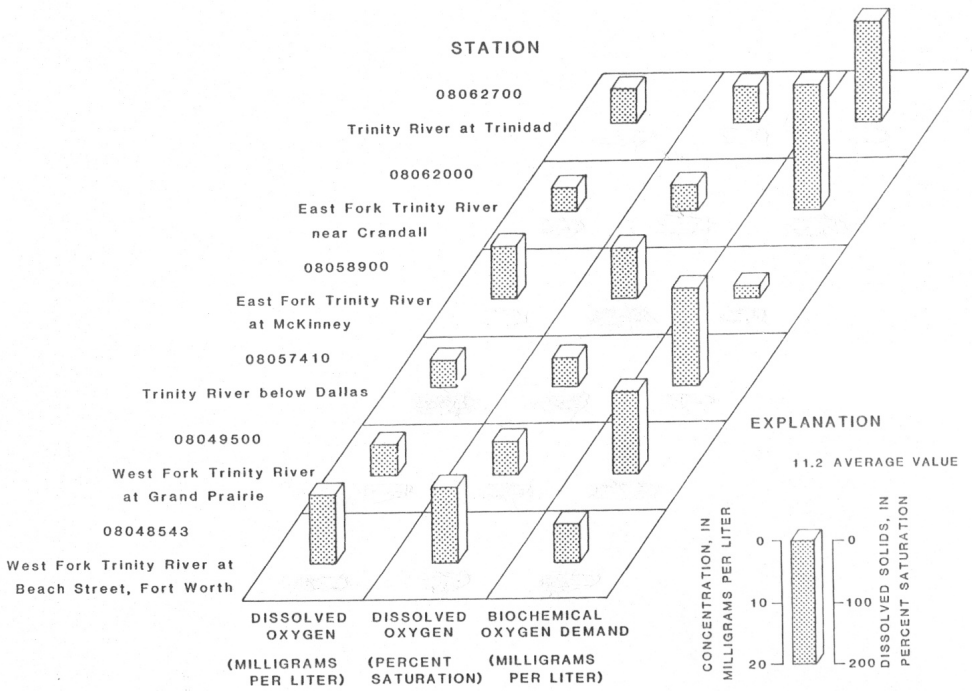


Figure 16.26 DO and BOD in the Trinity River, TX. From Wells et al., 1986.

16.4 Ice volumes of several glaciers at differing altitudes are compared in the following figure. How might it be changed to better show those differences?

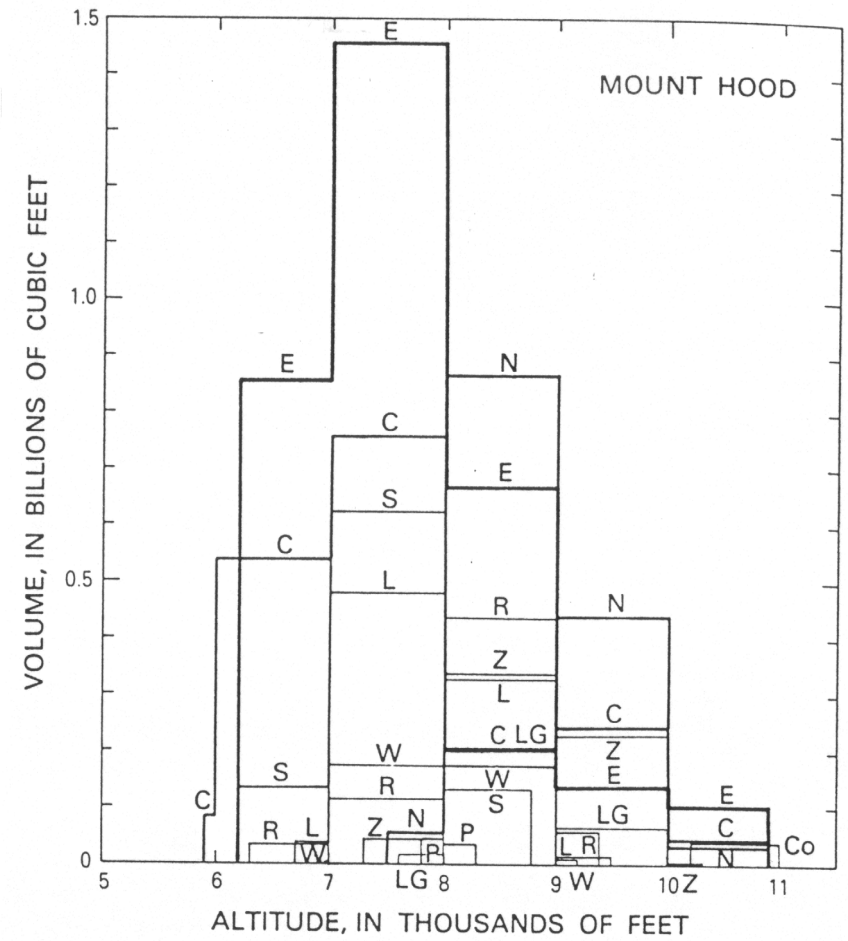


Figure 16.27 Ice volumes of Mount Hood glaciers. Each letter represents a different glacier. From Drieger and Kennard, 1986.

16.5 Water quality (major ions) was displayed for 13 numbered sites below. What other types of plots might have shown this more clearly?

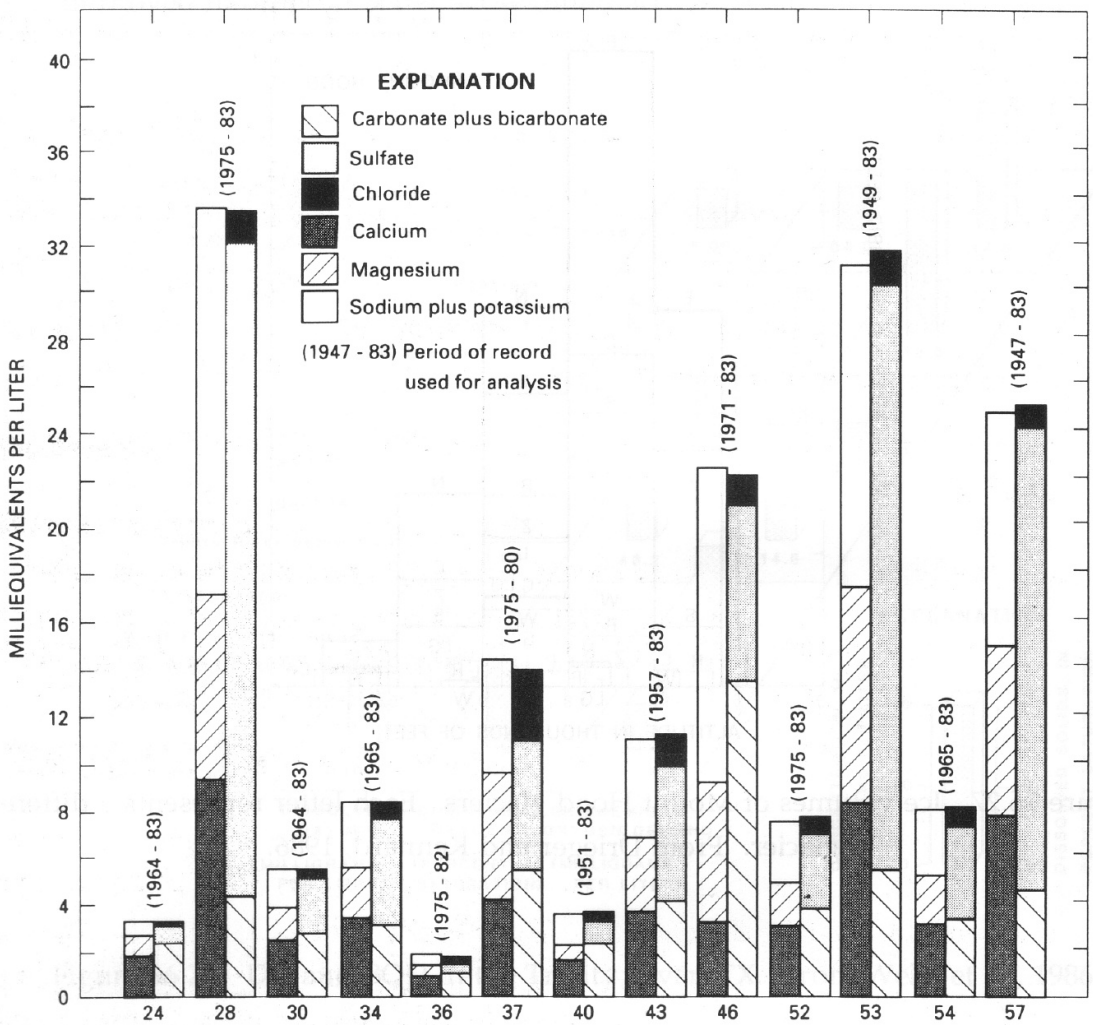


Figure 16.28 Chemical composition of streamwaters at 13 sites.
From Liebermann et al., 1989.

References Cited

Agresti, A., 1984, *Analysis of Ordinal Categorical Data*: Wiley & Sons, New York 287 p.

Aitchison, J. and Brown, J.A.C., 1981, *The Lognormal Distribution*: Cambridge Univ. Press, Cambridge, England, 176 p.

Alley, W. M., 1988, Using exogenous variables in testing for monotonic trends in hydrologic time series: *Water Resources Research 24*, 1955-1961.

Amemiya, T., 1981, Qualitative response models: a survey:*J. Economic Literature 19*, 1483-1536.

Anscombe, F. A., 1973, Graphs in Statistical Analysis: *Amer. Statistician. 27*, 17-21.

ASTM, 1983, Interlaboratory quality control procedures and a discussion on reporting low-level data: *Annual Book of ASTM Standards 11.01*, Chapter D, 4210-4283.

Atkinson, A. C., 1980, A note on the generalized information criterion for choice of a model: *Biometrika 67*, 413-418.

Bachman, L. J., 1984, Field and laboratory analyses of water from the Columbia Aquifer in Eastern Maryland: *Ground Water 22*, 460-467.

Beckman, R.J. and R. D. Cook, 1983, Outlier.....s: *Technometrics 25*, 119-149.

Bedinger, M. S., 1963, Relation between median grain size and permeability in the Arkansas River Valley, Arkansas: *U. S. Geological Survey Professional Paper 424-C*, 31-32.

Belsley, D. A.; E. Kuh; and R. E. Welsch, 1980, *Regression Diagnostics*: John Wiley, New York, 292 p.

Bhattacharyya, G.K., and R. A. Johnson, 1977, *Statistical Concepts and Methods*: John Wiley, New York, 639 p.

Blair, R. C. and J. J. Higgins, 1980, A comparison of the power of Wilcoxon's rank-sum statistic to that of student's t statistic under various nonnormal distributions: *J. Educ. Statistics 5* , 309-335.

Blom, G., 1958, *Statistical Estimates and Transformed Beta Variables*: John Wiley, New York, 68-75, 143-146.

Bloyd, R. M., P. B. Daddow, P. R. Jordan, and H. W. Lowham, 1986, Investigation of possible effects of surface coal mining on hydrology and landscape stability in part of the Powder River structural basin, Northeastern Wyoming: *U.S. Geological Survey Water-Resources Investigations Report 86-4329*, 101 p.

Boudette, E. L., F. C. Canney, J. E. Cotton, R. I. Davis, W. H. Ficklin, and J. M. Matooka, 1985, High levels of arsenic in the groundwaters of southeastern New Hampshire: *U.S. Geological Survey Open-File Report 85-202*, 23 p.

Box, G.E.P. and G. M., Jenkins, 1976, *Time Series Analysis: Forecasting and Control*: Holden Day, San Francisco, 575 p.

Bradley, J. V., 1968, *Distribution-Free Statistical Tests*: Prentice-Hall, Englewood Cliffs, NJ, 388 p.

Bradu, D., and Y. Mundlak, 1970, Estimation in lognormal linear models: *J. Amer. Statistical. Assoc., 65*, 198-211.

Bras, R.L. and Rodriguez-Iturbe, I., 1985, *Random Functions in Hydrology*: Addison-Wesley, Reading, MA, 559 p.

Breen, J. J.and P. E. Robinson, Eds., *Environmental Applications of Chemometrics*, ACS Symposium Series 292, Amer. Chemical Soc., Washington, D. C. 280 p.

Campbell, G. and J. H. Skillings, 1985, Nonparametric Stepwise Multiple Comparison Procedures: *J. Am. Stat. Assoc. 80*, 998-1003.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey, 1983, *Graphical Methods for Data Analysis*: PWS-Kent Publishing Co., Boston, 395 p.

Christensen, R., 1990, *Log-Linear Models*: Springer-Verlag, New York, NY, 408 p.

Cleveland, W.S., 1979, Robust Locally Weighted Regression and Smoothing Scatterplots: *J. Am. Stat. Assoc. 74*, 829-836.

Cleveland, W.S., 1984, Graphical methods for data presentation: full scale breaks, dot charts, and multibased logging: *American Statistician 38*, 270-280.

Cleveland, W.S., 1985, *The Elements of Graphing Data*: Wadsworth Books, Monterey, CA, 323 p.

Cleveland, W.S., and S. J. Devlin, 1988, Locally weighted regression: an approach to regression analysis by local fitting: *J. Am. Stat. Assoc. 83*, 596-610.

Cleveland, W.S., and R. McGill, 1983, A color-caused optical illusion on a statistical graph: *American Statistician 37*, 101-105.

Cleveland, W.S., and R. McGill, 1984a, Graphical perception: theory, experimentation, and application to the development of graphical methods: *J. Am. Stat. Assoc. 79*, 531-554.

Cleveland, W.S., and R. McGill, 1984b, The Many Faces of a Scatterplot: *J. Am. Stat. Assoc. 79*, 807-822.

Cleveland, W.S., and R. McGill, 1985, Graphical perception and graphical methods for analyzing scientific data: *Science 229*, 828-833.

Cohen, A.C., 1950, Estimating the mean and variance of normal populations for singly truncated and doubly truncated samples: *Annals of Mathematical Statistics 21*, 557-69.

Cohen, A. C., 1959, Simplified estimators for the normal distribution when samples are singly censored or truncated: *Technometrics 1*, 217-213.

Cohen, A.C., 1976, Progressively censored sampling in the three parameter log-normal distribution, *Technometrics 18*, 99-103.

Cohn, T. A., 1988, Adjusted maximum likelihood estimation of the moments of lognormal populations from type I censored samples: *U. S. Geological Survey Open-File Report 88-350,* 34 p.

Colby, B. R., C. H. Hembree, and F. H. Rainwater, 1956, Sedimentation and chemical quality of surface waters in the Wind River Basin, Wyoming: *U.S. Geological Survey Water-Supply Paper 1373*, 336 p.

Conover, W. J. and R. L. Iman, 1981, Rank transformation as a bridge between parametric and nonparametric statistics: *The Am. Stat. 35*, 3, 124-129.

Conover, W. L., 1980, *Practical Nonparametric Statistics, Second Edition*: John Wiley and Sons, New York, 493 p.

Conover, W. L., 1999, *Practical Nonparametric Statistics, Third Edition*: John Wiley and Sons, New York, 584 p.

Crabtree, R. W., I. D. Cluckie, and C. F. Forster, 1987, Percentile estimation for water quality data: *Water Research 21*, 583-590.

Crawford, C.G., J. R. Slack, and R. M. Hirsch, 1983, Nonparametric tests for trends in water-quality data using the Statistical Analysis System: *U.S. Geological Survey, Open-File Report 83-550*, 102 p.

Cunnane, C., 1978, Unbiased plotting positions - a review: *J. of Hydrology 37*, 205-222.

Davis, Robert E., and Gary D. Rogers, 1984, Assessment of selected ground-water-quality data in Montana: *U.S. Geological Survey Water-Resources Investigations Report 84-4173*, 177 p.

Dietz, E. J., 1985, The rank sum test in the linear logistic model: *American Statistician 39*, 322-325.

Donoho, Andrew, David L. Donoho, and Miriam Gasko, 1985, *MACSPIN Graphical Data Analysis Software*: D^2 Software, Inc., Austin, Texas, 185 p.

Doornkamp, J. C., and C. A. M. King, 1971, *Numerical Analysis in Geomorphology, An Introduction*: St. Martins Press, New York, NY, 372 p.

Draper, N. R. and Smith, H., 1981, *Applied Regression Analysis, Second Edition*: John Wiley and Sons, New York, 709 p.

Drieger, C. L., and P. M. Kennard, 1986, Ice volumes on Cascade volcanoes: Mount Ranier, Mount Hood, Three Sisters, and Mount Shasta: *U.S. Geological Survey Professional Paper 1365*, 28 p.

Duan, N., 1983, Smearing Estimate: A nonparametric retransformation method: *J. Am. Stat. Assoc., 78*, 605-610.

DuMouchel, W. H., and G. J. Duncan, 1983, Using sample survey weights in multiple regression analyses of stratified samples: *J. Am. Stat. Assoc., 78*, 535-543.

Durbin, J., and G. S. Watson, 1951, Testing for serial correlation in least squares regression, I and II: *Biometrika 37*, 409-428, and 38, 159-178.

Eckhardt, D.A., W.J. Flipse and E.T. Oaksford, 1989, Relation between land use and ground-water quality in the upper glacial aquifer in Nassau and Suffolk Counties, Long Island NY: *U.S. Geological Survey Water Resources Investigations Report 86-4142,* 26 p.

Everitt, B., 1978, *Graphical Techniques for Multivariate Data*: North-Holland Pubs, New York, 117 p.

Exner, M. E. and Spalding, R. F., 1976, Groundwater quality of the Central Platte Region, 1974: *Conservation and Survey Division Resource Atlas No. 2*, Institute of Agriculture and Natural Resources, University of Nebraska, Lincoln, Nebraska, 48 p.

Exner, M. E., 1985, Concentration of nitrate-nitrogen in groundwater, Central Platte Region, Nebraska, 1984: *Conservation and Survey Division*, Institute of Agriculture and Natural Resources, University of Nebraska, Lincoln, Nebraska, 1 p. (map).

Fawcett, R.F. and Salter, K.C., 1984, A Monte Carlo Study of the F Test and Three Tests Based on Ranks of Treatment Effects in Randomized Block Designs: *Communications in Statistics B13*, 213-225.

Fent, K., and J. Hunn, 1991, Phenyltins in water, sediment, and biota of freshwater marinas: *Environmental Science and Technology 25*, 956-963.

Ferguson, R. I. 1986, River loads underestimated by rating curves: *Water Resources Research 22*, 74-76.

Feth, J. H., C. E. Roberson, and W. L. Polzer, 1964, Sources of mineral constituents in water from granitic rocks, Sierra Nevada California and Nevada: *U.S. Geological Survey Water Supply Paper 1535-I*, 70 p.

Fisher, R. A., 1922, On the mathematical foundations of theoretical statistics, as quoted by Beckman and Cook, 1983, Outlier.....s: *Technometrics 25*, 119-149.

Foster, S. S. D., A. T. Ellis, M. Losilla-Penon, and J. V. Rodriguez-Estrada, 1985, Role of volcanic tuffs in ground-water regime of Valle Central, Costa Rica: *Ground Water 23*, 795-801.

Frenzel, S. A., 1988, Physical, chemical, and biological characteristics of the Boise River from Veterans Memorial Parkway, Boise to Star, Idaho, October 1987 to March 1988: *U.S. Geological Survey Water Resources Investigations 88-4206*, 48 p.

Frigge, M., D. C. Hoaglin, and B. Iglewicz, 1989, Some implementations of the boxplot: *American Statistician, 43*, 50-54.

Fusillo, T. V., J. J. Hochreiter, and D. G. Lord, 1985, Distribution of volatile organic compounds in a New Jersey coastal plain aquifer system, *Ground Water 23*, 354-360.

Gilbert, R. O., 1987, *Statistical Methods for Environmental Pollution Monitoring*: Van Nostrand Reinhold Co., New York, 320 p.

Gilliom, R. J., R. M. Hirsch, and E. J. Gilroy, 1984, Effect of censoring trace-level water-quality data on trend-detection capability: *Environ. Science and Technol., 18*, 530-535.

Gilliom, R. J., and D. R. Helsel, 1986, Estimation of distributional parameters for censored trace level water quality data, 1. Estimation techniques: *Water Resources Research 22,* 135-146.

Gleit, A., 1985, Estimation for small normal data sets with detection limits, *Environ. Science and Technol., 19*, 1201-1206.

Greek, B. F., 1987, Supply problems loom for big-volume adhesive monomers: *Chemical and Engineering News 65(16)*, April 20, 1987, p. 9.

Gringorten, I. I., 1963, A plotting rule for extreme probability paper: *J. of Geophysical Research 68*, 813-814.

Groggel, D.J. and J.H. Skillings, 1986, Distribution-Free Tests for Main Effects in Multifactor Designs: *The American Statistician, May, 40,* 99-102.

Groggel, D.J., 1987 A Monte Carlo Study of Rank Tests For Block Designs: *Communications in Statistics 16,* 601-620.

Grygier, J. C., J. R. Stedinger, and H.Yin, 1989, A Generalized Maintenance of Variance Extension Procedure for Extending Correlated Series: *Water Resources Research 25*, 345-349.

Haan, C. T., 1977. *Statistical Methods in Hydrology*: Iowa State University Press, Ames, Iowa**,** 378 p.

Hakanson, L., 1984, Sediment sampling in different aquatic environments: statistical aspects, *Water Resources Research 20*, 41-46.

Hald, A., 1949, Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point: *Skandinavisk Aktuarietidskrift 32*, 119-34.

Halfon, E., 1985, Regression method in ecotoxicology: A better formulation using the geometric mean functional regression: *Environ. Sci. and Technol. 19*, 747-749.

Hazen, A., 1914, Storage to be provided in the impounding reservoirs for municipal water supply: *Trans. Am. Soc. of Civil Engineers 77*, 1547-1550.

Helsel, D. R., 1983, Mine drainage and rock type influences on Eastern Ohio streamwater quality: *Water Resources Bulletin 19*, 881-887.

Helsel, D.R., 1987, Advantages of nonparametric procedures for analysis of water quality data: *Hydrological Sciences Journal 32,* 179-190.

Helsel, D.R. 1990, Less than obvious: Statistical treatment of data below the detection limit: *Environmental Science and Technology 24,* pp.1766-1774.

Helsel, D.R. 1992, Diamond in the rough: Enhancements to Piper diagrams: submitted to *Ground Water*.

Helsel, D.R. and T. A. Cohn, 1988, Estimation of descriptive statistics for multiply censored water quality data: *Water Resources Research 24,* 1997-2004.

Helsel, D.R. and R. J. Gilliom, 1986, Estimation of distributional parameters for censored trace level water quality data, 2. Verification and applications: *Water Resources Research 22,* 147-155.

Helsel, D.R. and R. M. Hirsch, 1988, Discussion of Applicability of the t-test for detecting trends in water quality variables: *Water Resources Bulletin 24,* 201-204.

Hem, J. D., 1985, Study and interpretation of the chemical characteristics of natural water: *U. S. Geological Survey Water Supply Paper 2254*, 263 p.

Henderson, Thomas, 1985, Geochemistry of ground-water in two sandstone aquifer systems in the Northern Great Plains in parts of Montana and Wyoming: *U.S. Geological Survey Professional Paper 1402-C*, 84 p.

Hirsch, R.M., 1982, A comparison of four record extension techniques: *Water Resources Research 15*, 1781-1790.

Hirsch, R. M., 1988, Statistical methods and sampling design for estimating step trends in surface-water quality: *Water Resources Bulletin, 24*, 493-503.

Hirsch, R.M. and J. R. Slack, 1984, A nonparametric trend test for seasonal data with serial dependence: *Water Resources Research 20*, 727-732.

Hirsch, R.M., J. R. Slack, and R. A. Smith, 1982. Techniques of trend analysis for monthly water quality data: *Water Resources Research 18*, 107-121.

Hirsch, R.M., and E. J. Gilroy, 1984, Methods of fitting a straight line to data: examples in water resources: *Water Resources Bulletin 20*, 705-711.

Hirsch, R.M., Alexander, and R. A. Smith, 1991, Selection of methods for the detection and estimation of trends in water quality: *Water Resources Research 27*, 803-813.

Hoaglin, D.C., 1983, Letter values: a set of order statistics: Chapter 2 in Hoaglin, D.C., F. Mosteller, and J.W. Tukey, eds., 1983. *Understanding Robust and Exploratory Data Analysis,* John Wiley, New York, NY, 447 p.

Hoaglin, D.C., F. Mosteller, and J.W. Tukey, eds., 1983. *Understanding Robust and Exploratory Data Analysis,* John Wiley, New York, NY, 447 p.

Hoaglin, David C., 1988, Transformations in everyday experience: *Chance 1*, 40-45.

Hodges, J.L., Jr. and E. L. Lehmann, 1963, Estimates of location based on rank tests, *Annals Mathematical Statistics 34*, 598-611.

Hoerl, A.E. and R. W. Kennard, 1970. Ridge regression: biased estimation for nonorthogonal problems: *Technometrics 12*, 55-67.

Hollander, M. and D. A. Wolfe, 1973, *Nonparametric Statistical Methods*: John Wiley and Sons, New York, 503 p.

Hollander, M. and D. A. Wolfe, 1999, *Nonparametric Statistical Methods, Second Edition*: John Wiley and Sons, New York, 787 p.

Holtschlag, D. J., 1987, Changes in water quality of Michigan streams near urban areas, 1973-84: *U.S. Geological Survey Water Resources Investigations 87-4035*, 120 p.

Hren, J., K. S. Wilson, and D. R. Helsel, 1984, A statistical approach to evaluate the relation of coal mining, land reclamation, and surface-water quality in Ohio: *U.S. Geological Survey Water Resources Investigations 84-4117*, 325 p.

Hull, L. C., 1984, Geochemistry of ground water in the Sacramento Valley, California: *U.S. Geological Survey Professional Paper 1401-B*, 36 p.

Iman, R. L., and J. M. Davenport, 1980, Approximations of the critical region of the Friedman statistic: *Communications in Statistics A, 9*, 571-595.

Iman, R. L., and W. J. Conover, 1983, *A Modern Approach to Statistics*: John Wiley and Sons, New York, 497 p.

Inman, D. L., 1952, Measures for describing the size distribution of sediments: *J. Sedimentary Petrology 22*, 125-145.

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood flow frequency: *Bulletin 17B of the Hydrology Subcommittee, U. S. Geological Survey*, Reston, VA., 185 p.

Janzer, V. J., 1986, Report of the U.S. Geological Survey's Analytical Evaluation Program -- Standard Reference Water Samples M6, M94, T95, N16, P8, and SED3: *Branch of Quality Assurance Report*, U. S. Geological Survey, Arvada, CO.

Jensen, A. L., 1973, Statistical analysis of biological data from preoperational-postoperational industrial water quality monitoring: *Water Research 7*, 1331-1347.

Johnson, N. M., G. F. Likens, F. H. Borman, D. W.Fisher, and R. S. Pierce, 1969, A working model for the variation in stream water chemistry at the Hubbard Brook Experimental Forest, New Hampshire: *Water Resources Research, 5,* 1353-1363.

Johnson, Richard A. and Dean W. Wichern, 1982, *Applied Multivariate Statistical Analysis*: Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 594 p.

Johnston, J., 1984, *Econometric Methods*: McGraw Hill, New York, 568 p.

Jordan, P. R., 1979, Relation of sediment yield to climatic and physical characteristics in the Missouri River basin: *U.S. Geological Survey Water-Resources Investigations 79-49*.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T. C. Lee, 1985, Qualitative and limited dependent variable models: Chap. 18 in *The Theory and Practice of Econometrics*: John Wiley and Sons, New York, 1019 p.

Junk, G. A., R.F. Spalding, and J.J. Richard, 1980, Areal, vertical, and temporal differences in ground-water chemistry: II. Organic constituents: *Journal of Environmental Quality 9*, 479-483.

Keith, L. H., Crummett, W., Deegan, J., Libby, R. A., Taylor, J. K., Wentler, G., 1983, Principles of environmental analysis: *Analytical Chemistry 55*, 2210-2218.

Kendall, M.G., 1938, A new measure of rank correlation: *Biometrika 30*, 81-93.

Kendall, M.G., 1975, *Rank Correlation Methods*, 4th edition: Charles Griffin, London. 202 p.

Kendall, M. G. and A. Stuart, *The Advanced Theory of Statistics*, vol. 2: Oxford University Press, New York, 748 p.

Kenney, J. F., and E. S. Keeping, 1954, *Mathematics of Statistics Part One*: D. Van Nostrand, New York, 102 p.

Kermack, K. A. and J. B. S. Haldane, 1950, Organic correlation and allometry: *Biometrika 37*, 30-41.

Kirby, W., 1974, Straight line fitting of an observation path by least normal squares: *U.S. Geological Survey Open-File Report 74-197*, 11p.

Kleiner, B., and T. E. Graedel, 1980, Exploratory data analysis in the geophysical sciences: *Reviews of Geophysics and Space Physics 18*, 699-717.

Knopman, D. S., 1990, Factors related to the water-yielding potential of rocks in the Piedmont and Valley and Ridge provinces of Pennsylvania: *U.S. Geological Survey Water-Resources Investigations Report 90-4174,* 52 p.

Kritskiy, S. N. and J. F. Menkel, 1968, Some statistical methods in the analysis of hydrologic data: *Soviet Hydrology Selected Papers 1*, 80-98.

Kruskal, W. H., 1953, On the uniqueness of the line of organic correlation, *Biometrics 9*, 47-58.

Kupper, L. L., and K. B. Hafner, 1989, How appropriate are popular sample size formulas?: *American Statistician 43*, 101-105.

Land, C. E., 1971, Confidence intervals for linear functions of the normal mean and variance: *Annals Mathematical Statistics 42*, 1187-1205.

Land, C. E., 1972, An evaluation of approximate confidence interval estimation methods for lognormal means: *Technometrics 14*, 145-158.

Langbein, W.B., 1960, Plotting positions in frequency analysis: *in* Dalrymple, T., Flood-frequency analysis: *U.S. Geological Survey Water-Supply Paper 1543-A,* p. 48-51.

Larsen, W.A. and S. J. McCleary, 1972. The use of partial residual plots in regression analysis: *Technometrics 14*, 781-790.

Latta, R., 1981, A monte carlo study of some two-sample rank tests with censored data: *Jour. American Statistical Association 76*, 713-719.

Lehmann, E.L., 1975. *Nonparametrics, Statistical Methods Based on Ranks*: Holden-Day, Oakland, CA, 457 p.

Lettenmaier, D.P., 1976, Detection of trends in water quality data from records with dependent observations: *Water Resources Research 12*, 1037-1046.

Lewandowsky, S. and I. Spence, 1989, Discriminating strata in scatterplots: *Jour. American Statistical Assoc., 84*, 682-688.

Liebermann, T., D. Mueller, J. Kircher, and A. Choquette, 1989, Characteristics and trends of streamflow and dissolved solids in the Upper Colorado River Basin: *U.S. Geological Survey Water-Supply Paper 2358*, 99 p.

Lin, S.D., and R. L. Evans, 1980, Coliforms and fecal streptococcus in the Illinois River at Peoria, 1971-1976: *Illinois State Water Survey Report of Investigations No. 93*, Urbana, IL, 28 p.

Lins, H., 1985, Interannual streamflow variability in the United States based on principal components: *Water Resources Research 21*, 691-701.

Looney, S. W., and T. R. Gulledge, 1985a, Use of the correlation coefficient with normal probability plots: *The American Statistician. 39*, 75-79.

Looney, S.W., and T. R. Gulledge, 1985b, Probability plotting positions and goodness of fit for the normal distribution: *The Statistician 34*, 297-303.

Maddala, G. S., 1983, *Limited-Dependent and Qualitative Variables in Econometrics*: Cambridge Univ. Press, Cambridge, U.K., 401 p.

Mann, H. B., 1945, Nonparametric test against trend: *Econometrica 13*, 245-259.

Marquardt, D.W., 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation: *Technometrics 12*, 591-612.

Martin, L., Leblanc, R. and N.K. Toan, 1993, Tables for the Friedman rank test: *Canadian Journal of Statistics 21, 1,* 39-43.

Matalas, N.C. and W. B. Langbein, 1962, Information content of the mean: *J.Geophysical. Res. 67*, 9, 3441-3448.

McCornack, R. L., 1965, Extended tables of the Wilcoxon matched pair signed rank statistic: *Journal of the American Statistical Association 60*, 864 – 871.

McCullagh, P., 1980, Regression Models for Ordinal Data: *Jour. Royal Stat. Soc. B, 42*, 109-142.

McGill, R., Tukey, J.W., and Larsen, W.A., 1978, Variations of box plots: *The American Statistician. 32*, 12-16.

McKean, J. W. and G. L. Sievers, Rank scores suitable for analyses of linear models under asymmetric error distributions:*Technometrics 31*, 207-218.

Meglen, R. R. and R. J. Sistko, 1985, Evaluating data quality in large data bases using pattern-recognition techniques: in Breen, J. J.and P. E. Robinson, Eds., *Environmental Applications of Chemometrics*, ACS Symposium Series 292, Amer. Chemical Soc., Washington, D. C. p. 16-33

Miesch, A., 1967, Methods of computation for estimating geochemical abundance: *U.S. Geological Survey Professional Paper 574-B*, 15 p.

Millard, S. P., and S. J. Deverel, Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits: *Water Resources Research 24,* 2087-98.

Miller, C. R., 1951, Analysis of flow-duration, sediment-rating curve method of computing sediment yield: *U. S. Bureau of Reclamation* unnumbered report, 65 p.

Miller, D. M., 1984, Reducing transformation bias in curve fitting: *The American Statistician. 38* , 124-126.

Miller, Timothy L., and Gonthier, Joseph B., 1984, Oregon ground-water quality and its relation to hydrogeological factors--a statistical approach: *U.S. Geological Survey Water Resources Investigations Report 84-4242*, 88 p.

Montgomery, D. C., 1991, *Introduction to Statistical Quality Control*: John Wiley, New York, NY, 674 p.

Montgomery, D. C., and E. A. Peck, 1982, *Introduction to Linear Regression Analysis*: John Wiley, New York, NY, 504 p.

Moody, D. W., E. W. Chase, and D. A. Aronson, compilers, 1986, National Water Summary 1985 -- Hydrologic events and surface-water resources: *U.S. Geological Survey Water-Supply Paper 2300*, p. 128.

Mosteller, F., and J. W. Tukey, 1977, *Data Analysis and Regression*: Addison-Wesley Publishers, Menlo Park, CA, 588 p.

Mustard, M.H., N.E. Driver, J. Chyr, and B.G. Hansen, 1987, U.S. Geological Survey urban stormwater data base of constituent storm loads: *U. S. Geological Survey Water Resources Investigation 87-4036,* 328 p.

Neter, J., W. Wasserman, and M. H. Kutner, 1985, *Applied Linear Statistical Models*, 2nd edition: Irwin Publishers, Homewood, Illinois, 1127 p.

Newman, M. C. and P. M. Dixon, 1990, UNCENSOR: A program to estimate means and standard deviations for data sets with below detection limit observations: *American Environmental Laboratory 4/90*, 26-30.

Noether, G. E., 1987, Sample size determination for some common nonparametric tests: *Journal American Statistical Assoc., 82*, 645-647.

Oltmann, R. N. and M. V. Shulters, 1989, Rainfall and Runoff Quantity and Quality Characteristics of Four Urban Land-Use Catchments in Fresno, California, October 1981 to April 1983: *U. S. Geological Survey Water-Supply Paper 2335,* 114 p.

Owen, W., and T. DeRouen, 1980, Estimation of the mean for lognormal data containing zeros and left-censored values, with applications to the measurement of worker exposure to air contaminants: *Biometrics 36,* 707-719.

Person, M., R. Antle, and D. B. Stephens, 1983, Evaluation of surface impoundment assessment in New Mexico: *Ground Water 21*, 679-688.

Piper, A. M., 1944, A graphic procedure in the geochemical interpretation of water analyses: *Transac. Amer. Geophysical Union 25*, p. 914-923.

Ponce, V. M., 1989, *Engineering Hydrology -- Principles and Practices*: Prentice-Hall, Englewood Cliffs, N.J., 640 p.

Porter, P. S., R. C. Ward, and H. F. Bell, 1988, The detection limit: *Environmental Science and Technol., 22*, 856-861.

Press, S. J. and S. Wilson, 1978, Choosing between logistic regression and discriminant analysis: *Journal American Statistical Assoc., 73*, 699-705.

Robertson, W. D., J. F. Barker, Y. LeBeau, and S. Marcoux, 1984, Contamination of an unconfined sand aquifer by waste pulp liquor: A case study: *Ground Water 22*, 191-197.

Rousseeuw, P. J., and A.M. Leroy, 1987, *Robust Regression and Outlier Detection*: John Wiley, New York, NY, 329 p.

Sanders, T. G., R. C. Ward, J.C. Loftis, T. D. Steele, D. D. Adrian, and V. Yevjevich, 1983, *Design of Networks for Monitoring Water Quality*: Water Resources Publications, Littleton, Colorado, 328 p.

SAS Institute, 1985, *SAS User's Guide: Statistics*: Version 5 Edition, SAS Institute Pub., Cary, NC, 470-476.

Sauer, V. B., W. O. Thomas, V. A. Stricker, and K. V. Wilson, 1983, Flood characteristics of urban watersheds in the United States: *U.S. Geological Survey Water-Supply Paper 2207*, 63 p.

Schertz, T.L., 1990, Trends in water quality data in Texas: *U.S. Geological Survey Water Resources Investigations Report 89-4178*, 177p.

Schertz, T.L., and R. M. Hirsch, 1985, Trend analysis of weekly acid rain data -- 1978-83: *U.S. Geological Survey Water Resources Investigations Report 85-4211*, 64 p.

Schmid, C. F., 1983, *Statistical Graphics*: John Wiley and Sons, New York, 212p.

Schroeder, L. J., M. H. Brooks, and T. C. Willoughby, 1987. Results of intercomparison studies for the measurement of pH and specific conductance at National Atmospheric Deposition Program/National Trends Network monitoring sites, October 1981-October 1985: *U.S. Geological Survey Water Resources Investigations Report 86-4363*, 22 p.

Sen, P.K., 1968. Estimates of regression coefficient based on Kendall's tau: *J. Am. Stat.Assoc., 63*, 1379-1389.

Shapiro, S. S., M. B. Wilk, and H. J. Chen, 1968, A comparative study of various tests for normality: *J. Amer. Stat. Assoc., 63*, 1343-1372.

Shapiro, S. S. and R. S. Francia, 1972, An approximate analysis of variance test for normality: *J. Amer. Stat. Assoc., 67*, 215-216.

Smith, R.A., R.B. Alexander, and M.G. Wolman, 1987, Analysis and interpretation of water quality trends in major U.S. rivers, 1974-81: *U.S. Geological Survey Water-Supply Paper 2307*, 25 p.

Sokal, R. R. and F. J. Rohlf, 1981, *Biometry*, *2nd Ed.:* W. H. Freeman and Co., San Francisco, 776 p.

Solley, W. B., E. B. Chase, and W.B. Mann, 1983, Estimated use of water in the United States in 1980: *U.S. Geological Survey Circular 1001*, 56 p.

Solley, W. B., C. F. Merck, and R. R. Pierce, 1988, Estimated use of water in the United States in 1985: *U.S. Geological Survey Circular 1004*, 82 p.

Sorenson, S. K., 1982, Water-quality management of the Merced River, California: *U.S. Geological Survey Open-File Report 82-450*, 46 p.

Stedinger, J. R., 1983, Confidence intervals for design events: *Journal of Hydraulic Eng., ASCE 109*, 13-27.

Stedinger, J. R., and G. D. Tasker, 1985, Regional hydrologic analysis: 1. Ordinary, weighted and generalized least squares compared: *Water Resources Research 21*, 1421-1432.

Stoline, M. R., 1981, The status of multiple comparisons: simultaneous estimation of all pairwise comparisons in one-way ANOVA designs: *The American Statistician, 35*, 134-141.

Tasker, G. D., 1980, Hydrologic regression with weighted least squares: *Water Resources Research 16*, 1107-1113.

Teissier, G., 1948, La relation d'allometrie sa signification statistique et biologique, *Biometrics 4*, 14-53.

Theil, H., 1950. A rank-invariant method of linear and polynomial regression analysis, 1, 2, and 3: *Ned. Akad. Wentsch Proc., 53*, 386-392, 521-525, and 1397-1412.

Travis, C.C., and M. L. Land, 1990, Estimating the mean of data sets with nondetectable values: *Environ. Sci. Technol. 24*, 961-962.

Tsiatis, A. A., 1980, A note on a goodness-of-fit test for the logistic regression model: *Biometrika 67*, 250-251.

Tufte, E. R., 1983, *The Visual Display of Quantitative Information*: Graphics Press, Cheshire, Connecticut., 197p.

Tukey, J. W., 1977, *Exploratory Data Analysis*: Addison-Wesley Pub., Reading MA, 506 p.

van Belle, G., and Hughes, J.P., 1984, Nonparametric tests for trend in water quality: *Water Resources Research 20*, 127-136.

Velleman, P. (1980), Definition and Comparison of Robust Nonlinear Data Smoothers, *J. Amer. Statistical. Assoc. 75*, 609-615.

Velleman, P. F. and D. C. Hoaglin, 1981, *Applications, Basics, and Computing of Exploratory Data Analysis*: Duxbury Press, Boston, MA, 354 p.

Vogel, R. M., 1986, The probability plot correlation coefficient test for the normal, lognormal, and gumbel distributional hypotheses: *Water Res. Research 22*, 587-590.

Vogel, Richard M., and Charles N. Kroll, 1989, Low-flow frequency analysis using probability-plot correlation coefficients: *J. of Water Resources Planning and Management, ASCE, 115*, 338-357.

Vogel, R.M. and J.R. Stedinger, 1985, Minimum variance streamflow record augmentation procedures: *Water Resources Research 21*, 715-723.

Walker, S. H., and D. B. Duncan, 1967, Estimation of the probability of an event as a function of several independent variables: *Biometrika 54*, 167-179.

Walpole, R. E. and R. H. Myers, 1985, *Probability and Statistics for Engineers and Scientists,* 3rd ed.: MacMillan Pub., New York, 639 p.

Walton, W. C., 1970, *Groundwater Resource Evaluation*: McGraw-Hill, New York, 664 p.

Weibull, W., 1939, *The Phenomenon of Rupture in Solids*: Ingeniors Vetenskaps Akademien Handlinga 153, Stockholm, p. 17.

Welch, A. H., M. S. Lico, and J. L. Hughes, 1988, Arsenic in ground water of the western United States: *Ground Water 26*, 333-338.

Wells, F. C., J. Rawson, and W. J. Shelby, 1986, Areal and temporal variations in the quality of surface water in hydrologic accounting unit 120301, Upper Trinity River basin, Texas: *U.S. Geological Survey Water-Resources Investigations Report 85-4318*, 135 p.

Wilcoxon, F., 1945, Individual comparisons by ranking methods: *Biometrics, 1*, 80-83.

Wilk, M. B. and H. J. Chen, 1968, A comparative study of various tests for normality: *J. Amer. Stat. Assoc. 63*, 1343-72.

Williams, G. P., and Wolman, M.G., 1984, Downstream effects of dams on alluvial rivers*: U.S. Geological Survey Professional Paper 1286*, 83 p.

Wright, W. G., 1985, Effects of fracturing on well yields in the coal field areas of Wise and Dickenson Counties, Southwestern Virginia: *U.S. Geological Survey Water-Resources Investigations Report 85-4061*, 21 p.

Xhoffer, C., P. Bernard, and R. Van Grieken, Chemical characterization and source apportionment of individual aerosol particles over the North Sea and the English Channel using multivariate techniques: *Environmental Science and Technology 24*, 1470-1478.

Yee, J. J. S., and C. J. Ewart, 1986, Biological, morphological, and chemical characteristics of Wailuku River, Hawaii: *U.S. Geological Survey Water Resources Investigations Report 86-4043*, 69 p.

Yorke, T. H., J. K. Stamer, and G. L. Pederson, G.L., Effects of low-level dams on the distribution of sediment, trace metals, and organic substances in the Lower Schuylkill River Basin, Pennsylvania: *U.S. Geological Survey Water-Supply Paper 2256-B*, 53 p.

Zar, Jerrold H., 1999, *Biostatistical Analysis, Fourth Edition*. Prentice-Hall, Saddle River, New Jersey, 663 p.

Zelen, N., and N. C. Severo, 1964, Probability Functions: Chapter 26 in Abramowitz, M. and I. A. Stegun, eds., *Handbook of Mathematical Functions*: U.S. National Bureau of Standards, Applied Mathematics Series No. 55, Wash, D.C., 1045 p.

Appendix A

Construction of Boxplots

The upper and lower limits of the central box are defined using either quartiles or hinges. These definitions are clarified below. Then the influence of each definition on the position of the whiskers is demonstrated. Definitions used by commercial software packages are listed, including one non-conventional form called a "box graph".

Quartiles

Quartiles are the 25th, 50th and 75th percentiles of a data set, as defined in chapter 1. Consider a data set X_i, i=1,...n. Computation of percentiles follows the equation

$$p_j = X_{(n+1)\bullet j}$$

where n is the sample size of X_i,

j is the fraction of data less than or equal to the percentile value (for the 3 quartiles, j= .25, .50, and .75).

Non-integer values of $(n+1)\bullet j$ imply linear interpolation between adjacent values of X. Computation of quartiles for two small example data sets is illustrated in Table 1.

Hinges

Tukey (1977) used values for the ends of the box which, along with the median, divided the data into four equal parts. These "fourths" or "hinges" are defined as:

Lower hinge h_L = median of all observations less than or equal to the sample median.

Upper hinge h_U = median of all observations equal to or greater than the overall sample median.

They may also be defined as:

$$\text{Lower hinge } h_L = X_L, \text{ where } L = \frac{\text{integer}\,[\,(n+3)/2\,]}{2}, \text{ and}$$

$$\text{Upper hinge } h_U = X_U, \text{ where } U = (n+1) - L.$$

where "integer [ ]" is the integer portion of the number in brackets. For example, integer [5.7] = 5. Again, non-integer values of L and U imply interpolation. With hinges, however, this will always

be halfway between adjacent data points. Therefore, hinges are always either data values themselves, or averages of two data points, and so are easier to compute by hand than are percentiles. Hinges will generally be similar to quartiles for large (n> 30) sample sizes. For smaller data sets, differences will be more apparent. For example, when n=12 the lower hinge is halfway between the 3rd and 4th data points, while the lower quartile is one-quarter of the way between the two points (see Table 1) . Both measures split the data into one-fourth below and three fourths above their value. Either are acceptable for use in boxplots.

Table A1

A. For the following data X_i of sample size n=11:

2 3 5 45 46 47 48 50 90 151 208

$p_{.25}$ =	lower quartile =	$X_{(n+1)\bullet.25}$	$=X_3$	= 5.
$p_{.75}$ =	upper quartile =	$X_{(n+1)\bullet.75}$	$= X_9$	= 90.
$p_{.50}$ =	median =	$X_{(n+1)\bullet.50}$	$= X_6$	= 47.
h_l =	lower hinge =	median [2 3 5 45 46 47]		= 25.
h_u =	upper hinge =	median [47 48 50 90 151 208]		= 70.

B. For sample size n=12, and data X_i, i=1,...n equal to:

2 3 5 45 46 47 48 49 50 90 151 208

$p_{.25}$ =	lower quartile =	$X_{(n+1)\bullet.25}$	$= X_{3.25} = X_3 + 0.25\bullet(X_4 - X_3)$	= 15.
$p_{.75}$ =	upper quartile =	$X_{(n+1)\bullet.75}$	$= X_{9.75} = X_9 + 0.75\bullet(X_{10} - X_9)$	= 80.
$p_{.50}$ =	median =	$X_{(n+1)\bullet.50}$	$= X_{6.5} = X_6 + 0.50\bullet(X_7 - X_6)$	= 47.5.
h_l =	lower hinge =	median [2 3 5 45 46 47]		= 25.
h_u =	upper hinge =	median [48 49 50 90 151 208]		= 70.

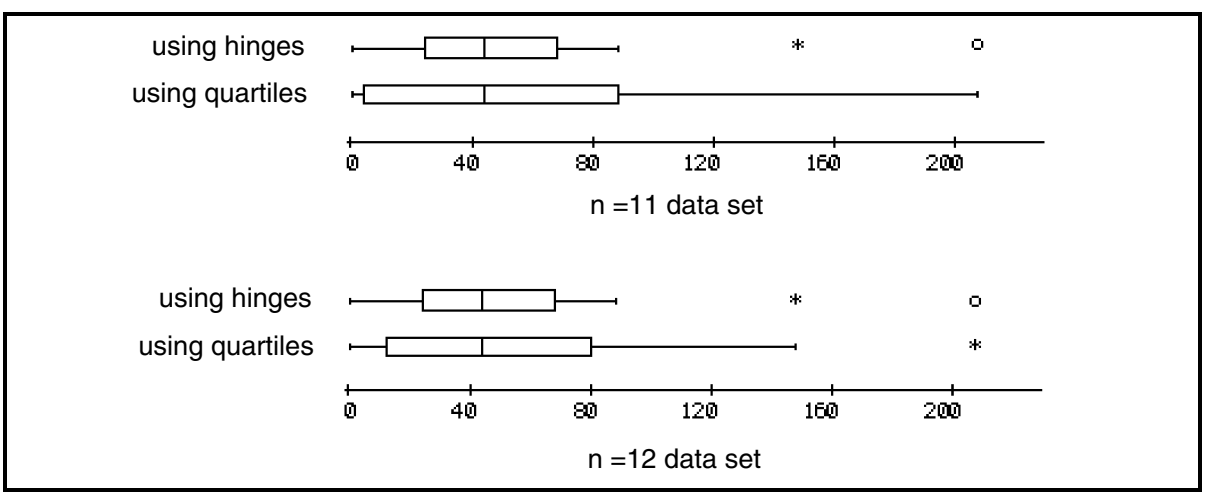


Figure A1. Boxplots for the Table A1 data

Figure A1 shows standard boxplots for the Table 1 data using both percentiles and hinges. Data in Table 1 were designed to maximize differences between the two measures. Real data, and larger sample sizes, will evidence much smaller differences. Note that the definitions of the box boundaries directly affect whisker lengths, and also determines which data are plotted as "outside" values.

It would be ideal if all software used the same conventions for drawing boxplots. However, that has not happened. Software written by developers who stick to the original definitions prefer hinges; those who want box boundaries to agree with tabled percentiles use quartiles. The Table 1 data can be used to determine which convention is used to produce boxplots.

Non-conventional definitions

Other statistical software use another (non-conventional) value for the box boundaries (Frigge and others, 1989). They use the next highest data value for the lower box boundary whenever n/4 is not an integer. This avoids all interpolation. Note that n, not n+1, is used.

StatView uses a percentile-type boxplot similar to the truncated boxplot, except that the upper and lower 10 percent of data are plotted as individual points. The weakness of this scheme is that 10 percent of the data will always be plotted individually at each end of the plot, and so it is less effective for defining and emphasizing unusual values. Also important is that StatView uses yet another definition for the box boundaries, $X_{(n+2)\bullet j}$, in calculating the quartiles. This non-conventional boxplot was called a "box graph" by Cleveland (1985).

Therefore some statistical software will produce boxes differing from conventional boxplots, particularly for small data sets.

Boxplots for Censored Data

Data sets whose values include some observations known only to be below (or above) a limit or threshold can also be effectively displayed by boxplots. First set all values below the threshold to some value less than (not equal to) the reporting limit. The actual value is not important, and could be 0, one-half the reporting limit, etc. Produce the boxplot. Then draw a line across the graph at the value of the threshold, and erase all lines below this value from the graph.

This procedure was used for data in figure A2. If less than 25 percent of the data are below the threshold, this procedure will affect at most only the lower whisker (as in the Hoover Dam through Morelos Dam boxplots). If between 25 and 75 percent are below the threshold, the box will be partially hidden below the threshold (as in the CO-UT Line and Cisco boxes). If more than 75 percent of the data are below the threshold, part of the upper whisker and outside values will be visible above the threshold, as in the Lees Ferry box. In each case, these boxplots accurately and fairly illustrate both the distribution of data above the threshold, and the percentage of data below the threshold.

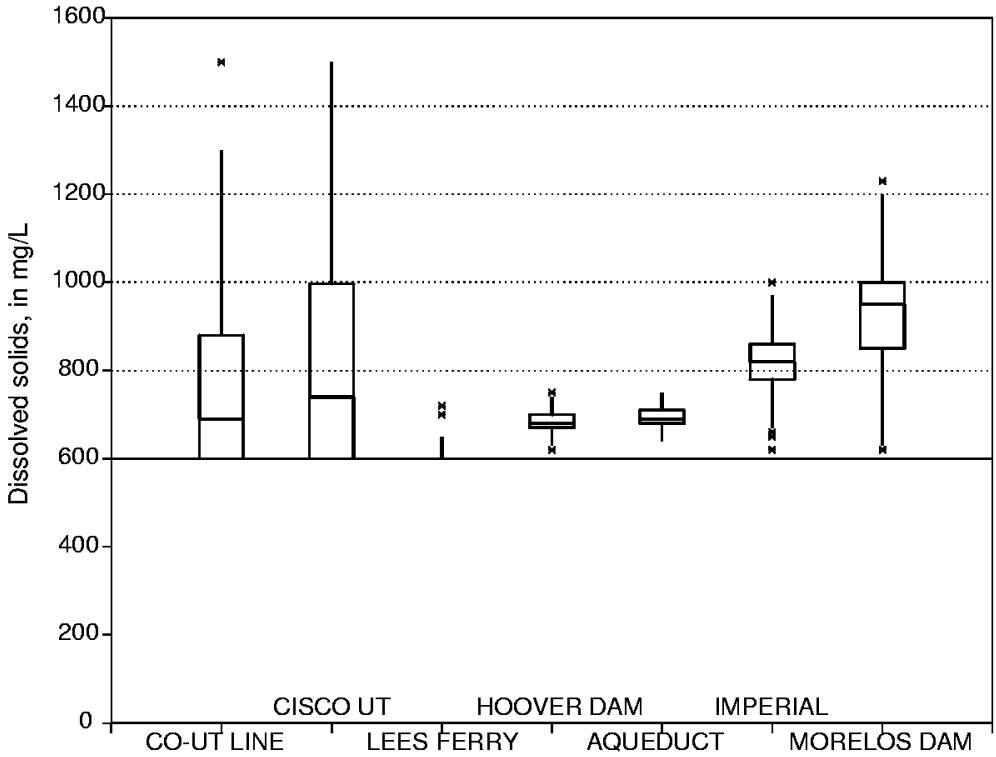


Figure A2. Dissolved solids concentrations along the Colorado River, artificially censored at a threshold of 600 mg/L.

A second alternative for boxplots of censored data is to estimate the percentiles falling below the threshold, and drawing dashed portions of the box below the threshold using these estimates. Helsel and Cohn (1988) have compared methods for estimating these percentiles. When multiple thresholds occur, such as thresholds which have changed over time or between laboratories, a solid line can be drawn across the plot at the highest threshold. Portions of the boxes above the highest threshold will be correct as long as each censored observation is assigned some value below its threshold. Quartiles falling below the highest threshold should be determined by using the methods recommended by Helsel and Cohn (1988). All lines below the highest threshold are only estimates, and should be drawn as dashed lines on the plot.

Displaying confidence intervals

As an aid for displaying whether two groups of data have different medians, confidence intervals for the median as defined in chapter 3 can be added to boxplots. When boxplots are placed side by side, their medians are significantly different if the confidence intervals do not overlap. Three methods of displaying these intervals are shown in figure A3. In the first method (A), the box is "notched" at both upper and lower limits, making the box narrower for all values within the interval. In the second (B), parentheses are drawn within the box at each limit. Shading is used in (C) to illustrate interval width. If displaying differences in medians is not of primary interest, these

methods add visual confusion to boxplots and are probably best avoided. Confusion is compounded when the interval width falls beyond the 25th or 75th percentiles. Of the three, shading seems the easiest to visualize and least confusing.

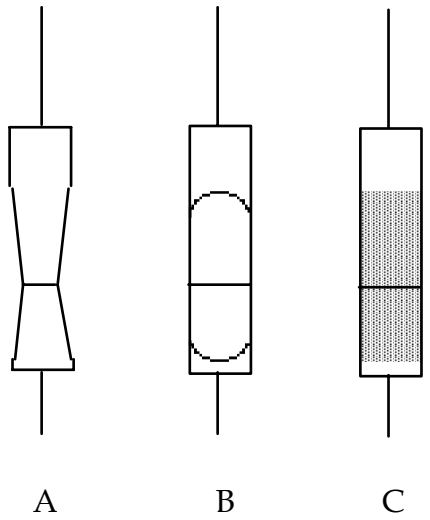


Figure A3. Methods for displaying confidence interval of median using a boxplot.
A. Notched boxplots B. Parentheses C. Shaded boxplot

Appendix B

Tables

Table B1 Cunnane plotting positions for n = 1 to 20

Table B2 Normal quantiles for Cunnane plotting positions of Table B1

Table B3 Critical values for the PPCC test for normality

Table B4 Quantiles (p-values) for the rank-sum test

Table B5 Quantiles (p-values) for the sign test

Table B6 Critical test statistic values for the signed-rank test

Table B7 Critical test statistic values for the Friedman test

Table B8 Quantiles (p-values) for Kendall's tau (τ)

Table B1. Cunnane plotting positions for sample sizes n = 1 to 20

									i										
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		N= 5																	
.12	.31	.50	.69	.88															
		N= 6																	
.10	.26	.42	.58	.74	.90														
		N= 7																	
.08	.22	.36	.50	.64	.78	.92													
		N= 8																	
.07	.20	.32	.44	.56	.68	.80	.93												
		N= 9																	
.07	.17	.28	.39	.50	.61	.72	.83	.93											
		N= 10																	
.06	.16	.25	.35	.45	.55	.65	.75	.84	.94										
		N= 11																	
.05	.14	.23	.32	.41	.50	.59	.68	.77	.86	.95									
		N= 12																	
.05	.13	.21	.30	.38	.46	.54	.62	.70	.79	.87	.95								
		N= 13																	
.05	.12	.20	.27	.35	.42	.50	.58	.65	.73	.80	.88	.95							
		N= 14																	
.04	.11	.18	.25	.32	.39	.46	.54	.61	.68	.75	.82	.89	.96						
		N= 15																	
.04	.11	.17	.24	.30	.37	.43	.50	.57	.63	.70	.76	.83	.89	.96					
		N= 16																	
.04	.10	.16	.22	.28	.35	.41	.47	.53	.59	.65	.72	.78	.84	.90	.96				
		N= 17																	
.03	.09	.15	.21	.27	.33	.38	.44	.50	.56	.62	.67	.73	.79	.85	.91	.97			
		N= 18																	
.03	.09	.14	.20	.25	.31	.36	.42	.47	.53	.58	.64	.69	.75	.80	.86	.91	.97		
		N- 19																	
.03	.08	.14	.19	.24	.29	.34	.40	.45	.50	.55	.60	.66	.71	.76	.81	.86	.92	.97	
		N= 20																	
.03	.08	.13	.18	.23	.28	.33	.38	.43	.48	.52	.57	.62	.67	.72	.77	.82	.87	.92	.97

Table B2. Upper tail normal quantiles for the plotting positions of Table B1

(for lower tail quantiles, multiply all nonzero quantiles by −1)

	N= 5								
0.000	0.502	1.198							
	N= 6								
0.203	0.649	1.300							
	N= 7								
0.000	0.355	0.765	1.383						
	N= 8								
0.153	0.475	0.859	1.453						
	N= 9								
0.000	0.276	0.575	0.939	1.513					
	N= 10								
0.123	0.377	0.659	1.007	1.565					
	N= 11								
0.000	0.225	0.463	0.732	1.067	1.611				
	N= 12								
0.103	0.313	0.538	0.796	1.121	1.653				
	N= 13								
0.000	0.191	0.389	0.604	0.852	1.169	1.691			
	N= 14								
0.088	0.267	0.456	0.663	0.904	1.212	1.725			
	N= 15								
0.000	0.165	0.336	0.517	0.716	0.950	1.252	1.757		
	N= 16								
0.077	0.234	0.397	0.571	0.765	0.992	1.289	1.787		
	N= 17								
0.000	0.146	0.295	0.452	0.620	0.809	1.031	1.323	1.814	
	N= 18								
0.069	0.208	0.351	0.502	0.666	0.849	1.067	1.354	1.839	
	N= 19								
0.000	0.131	0.264	0.402	0.548	0.707	0.887	1.101	1.383	1.864
	N= 20								
0.062	0.187	0.315	0.449	0.591	0.746	0.922	1.133	1.411	1.886

Table B3. Critical r* values for the probability plot correlation coefficient test of normality (from Looney and Gulledge, 1985a)

© American Statistical Association. Used with permission.

[reject H_0: data are normal when PPCC $r < r^*$]

	α-level					
n	**.005**	**.010**	**.025**	**.050**	**.100**	**.250**
3	.867	.869	.872	.879	.891	.924
4	.813	.824	.846	.868	.894	.931
5	.807	.826	.856	.880	.903	.934
6	.820	.838	.866	.888	.910	.939
7	.828	.850	.877	.898	.918	.944
8	.840	.861	.887	.906	.924	.948
9	.854	.871	.894	.912	.930	.952
10	.862	.879	.901	.918	.934	.954
11	.870	.886	.907	.923	.938	.957
12	.876	.892	.912	.928	.942	.960
13	.885	.899	.918	.932	.945	.962
14	.890	.905	.923	.935	.948	.964
15	.896	.910	.927	.939	.951	.965
16	.899	.913	.929	.941	.953	.967
17	.905	.917	.932	.944	.954	.968
18	.908	.920	.935	.946	.957	.970
19	.914	.924	.938	.949	.958	.971
20	.916	.926	.940	.951	.960	.972
21	.918	.930	.943	.952	.961	.973
22	.923	.933	.945	.954	.963	.974
23	.925	.935	.947	.956	.964	.975
24	.927	.937	.949	.957	.965	.976
25	.929	.939	.951	.959	.966	.976
26	.932	.941	.952	.960	.967	.977
27	.934	.943	.953	.961	.968	.978
28	.936	.944	.955	.962	.969	.978
29	.939	.946	.956	.963	.970	.979
30	.939	.947	.957	.964	.971	.979
31	.942	.950	.958	.965	.972	.980
32	.943	.950	.959	.966	.972	.980
33	.944	.951	.961	.967	.973	.981
34	.946	.953	.962	.968	.974	.981
35	.947	.954	.962	.969	.974	.982
36	.948	.955	.963	.969	.975	.982
37	.950	.956	.964	.970	.976	.983
38	.951	.957	.965	.971	.976	.983
39	.951	.958	.966	.971	.977	.983
40	.953	.959	.966	.972	.977	.984

Table B3. Cont.

	α-level					
n	**.005**	**.010**	**.025**	**.050**	**.100**	**.250**
41	.953	.960	.967	.973	.977	.984
42	.954	.961	.968	.973	.978	.984
43	.956	.961	.968	.974	.978	.984
44	.957	.962	.969	.974	.979	.985
45	.957	.963	.969	.974	.979	.985
46	.958	.963	.970	.975	.980	.985
47	.959	.965	.971	.976	.980	.986
48	.959	.965	.971	.976	.980	.986
49	.961	.966	.972	.976	.981	.986
50	.961	.966	.972	.977	.981	.986
55	.965	.969	.974	.979	.982	.987
60	.967	.971	.976	.980	.984	.988
65	.969	.973	.978	.981	.985	.989
70	.971	.975	.979	.983	.986	.990
75	.973	.976	.981	.984	.987	.990
80	.975	.978	.982	.985	.987	.991
85	.976	.979	.983	.985	.988	.991
90	.977	.980	.984	.986	.988	.992
95	.979	.981	.984	.987	.989	.992
100	.979	.982	.985	.987	.989	.992

Table B4. Quantiles (p-values) for the rank-sum test statistic W_{rs}

$p = \text{Prob}\,[W_{rs} \geq x] = \text{Prob}\,[W_{rs} \leq x^*]$

n [smaller sample size] = 3

m=4			m=5			m=6			m=7			m=8			m=9			m=10		
x	p	x*	x	p	x*	x	p	x*	x	p	x*	x	p	x*	x	p	x*	x	p	x*
16	.114	8	18	.125	9	20	.131	10	22	.133	11	24	.139	12	27	.105	12	29	.108	13
17	.057	7	19	.071	8	21	.083	9	23	.092	10	25	.097	11	28	.073	11	30	.080	12
18	.029	6	20	.036	7	22	.048	8	24	.058	9	26	.067	10	29	.050	10	31	.056	11
19	0	5	21	.018	6	23	.024	7	25	.033	8	27	.042	9	30	.032	9	32	.038	10
			22	0	5	24	.012	6	26	.017	7	28	.024	8	31	.018	8	33	.024	9
						25	0	5	27	.008	6	29	.012	7	32	.009	7	34	.014	8
									28	0	5	30	.006	6	33	.005	6	35	.007	7

n [smaller sample size] = 4

m=4			m=5			m=6			m=7			m=8			m=9			m=10		
x	p	x*	x	p	x*	x	p	x*	x	p	x*	x	p	x*	x	p	x*	x	p	x*
22	.171	14	25	.143	15	28	.129	16	31	.115	17	34	.107	18	36	.130	20	39	.120	21
23	.100	13	26	.095	14	29	.086	15	32	.082	16	35	.077	17	37	.099	19	40	.094	20
24	.057	12	27	.056	13	30	.057	14	33	.055	15	36	.055	16	38	.074	18	41	.071	19
25	.029	11	28	.032	12	31	.033	13	34	.036	14	37	.036	15	39	.053	17	42	.053	18
26	.014	10	29	.016	11	32	.019	12	35	.021	13	38	.024	14	40	.038	16	43	.038	17
27	0	9	30	.008	10	33	.010	11	36	.012	12	39	.014	13	41	.025	15	44	.027	16
			31	0	9	34	.005	10	37	.006	11	40	.008	12	42	.017	14	45	.018	15
									38	.003	10	41	.004	11	43	.010	13	46	.012	14
															44	.006	12	47	.007	13

n [smaller sample size] = 5

m=5			m=6			m=7			m=8			m=9			m=10		
x	p	x*	x	p	x*	x	p	x*	x	p	x*	x	p	x*	x	p	x*
34	.111	21	37	.123	23	41	.101	24	44	.111	26	47	.120	28	51	.103	29
35	.075	20	38	.089	22	42	.074	23	45	.085	25	48	.095	27	52	.082	28
36	.048	19	39	.063	21	43	.053	22	46	.064	24	49	.073	26	53	.065	27
37	.028	18	40	.041	20	44	.037	21	47	.047	23	50	.056	25	54	.050	26
38	.016	17	41	.026	19	45	.024	20	48	.033	22	51	.041	24	55	.038	25
39	.008	16	42	.015	18	46	.015	19	49	.023	21	52	.030	23	56	.028	24
40	.004	15	43	.009	17	47	.009	18	50	.015	20	53	.021	22	57	.020	23
			44	.004	18	48	.005	17	51	.009	19	54	.014	21	58	.014	22
									52	.005	18	55	.009	20	59	.010	21
												56	.006	19	60	.006	20

n [smaller sample size] = 6

m=6			m=7			m=8			m=9			m=10		
x	p	x*	x	p	x*	x	p	x*	x	p	x*	x	p	x*
47	.120	31	51	.117	33	55	.114	35	59	.112	37	63	.110	39
48	.090	30	52	.090	32	56	.091	34	60	.091	36	64	.090	38
49	.066	29	53	.069	31	57	.071	33	61	.072	35	65	.074	37
50	.047	28	54	.051	30	58	.054	32	62	.057	34	66	.059	36
51	.032	27	55	.037	29	59	.041	31	63	.044	33	67	.047	35
52	.021	26	56	.026	28	60	.030	30	64	.033	32	68	.036	34
53	.013	25	57	.017	27	61	.021	29	65	.025	31	69	.028	33
54	.008	24	58	.011	26	62	.015	28	66	.018	30	70	.021	32
			59	.007	25	63	.010	27	67	.013	29	71	.016	31
									68	.009	28	72	.011	30
												73	.008	29

TABLE B4 continued

n [smaller sample size] = 7

m=7			m=8			m=9			m=10		
x	p	x*	x	p	x*	x	p	x*	x	p	x*
61	.159	44	65	.168	47	70	.150	49	74	.15752	
62	.130	43	66	.140	46	71	.126	48	75	.13551	
63	.104	42	67	.116	45	72	.105	47	76	.11550	
64	.082	41	68	.095	44	73	.087	46	77	.09749	
65	.064	40	69	.076	43	74	.071	45	78	.08148	
66	.049	39	70	.060	42	75	.057	44	79	.06747	
67	.036	38	71	.047	41	76	.045	43	80	.05446	
68	.027	37	72	.036	40	77	.036	42	81	.04445	
69	.019	36	73	.027	39	78	.027	41	82	.03544	
70	.013	35	74	.020	38	79	.021	40	83	.02843	
71	.009	34	75	.014	37	80	.016	39	84	.02242	
72	.006	33	76	.010	36	81	.011	38	85	.01741	
			77	.007	35	82	.008	37	86	.01240	
						83	.006	36	87	.00939	

n [smaller sample size] = 9

m=9			m=10		
x	p	x*	x	p	x*
98	.149	73	104	.13976	
99	.129	72	105	.12175	
100	.111	71	106	.10674	
101	.095	70	107	.09173	
102	.081	69	108	.07872	
103	.068	68	109	.06771	
104	.057	67	110	.05670	
105	.047	66	111	.04769	
106	.039	65	112	.03968	
107	.031	64	113	.03367	
108	.025	63	114	.02766	
109	.020	62	115	.02265	
110	.016	61	116	.01764	
111	.012	60	117	.01463	
112	.009	59	118	.01162	
113	.007	58	119	.00961	

n [smaller sample size] = 8

m=8			m=9			m=10		
x	p	x*	x	p	x*	x	p	x*
79	.13957		84	.138	60	89	.137	63
80	.11756		85	.118	59	90	.118	62
81	.09755		86	.100	58	91	.102	61
82	.08054		87	.084	57	92	.086	60
83	.06553		88	.069	56	93	.073	59
84	.05252		89	.057	55	94	.061	58
85	.04151		90	.046	54	95	.051	57
86	.03250		91	.037	53	96	.042	56
87	.02549		92	.030	52	97	.034	55
88	.01948		93	.023	51	98	.027	54
89	.01447		84	.018	50	99	.022	53
90	.01046		95	.014	49	100	.017	52
91	.00745		96	.010	48	101	.013	51
92	.00544		97	.008	47	102	.010	50

n [smaller sample size] = 10

m=10		
x	p	x*
119	.157	91
120	.140	90
121	.124	89
122	.109	88
123	.095	87
124	.083	86
125	.072	85
126	.062	84
127	.053	83
128	.045	82
129	.038	81
130	.032	80
131	.026	79
132	.022	78
133	.018	77
134	.014	76
135	.012	75
136	.009	74
137	.007	73
138	.006	72

Table generated by D. Helsel

Table B5. -- Quantiles (p-values) for the sign test statistic S^+

Quantiles for the sign test are identical to quantiles of the binomial distribution with percentile p=0.5. The approximation given in chapter 6 and used by most statistical software packages can be used for n ≥ 20. Statistics textbooks that contain a table of exact quantiles for the binomial distribution for sizes below 20 include Hollander and Wolfe (1999) and Zar (1999).

An online table of exact quantiles for the binomial distribution can be found as of 5/2002 at: http://faculty.vassar.edu/lowry/binomial01.html

An example of using this online table:
Enter n (the number of data pairs) and p (=0.5). An exact table will be printed. P-values are cumulative probabilities, or values of the cumulative distribution function (cdf). For small values of the test statistic S^+ (called k in the online table) – values below n/2, use the "Down" column to read off a one-sided p-value for the sign test. For S^+ larger than n/2, use the "Up" column. The example output below is for n=13. A one-sided p-value for $S^+ = 4$ (the probability of getting an $S^+ \leq 4$) is 0.133. The p-value for $S^+ = 9$ (the probability of getting an $S^+ \geq 9$) also equals 0.133. For a two-sided test, p = 0.266.

		Cumulative Probability	
k	**Exact Probability**	**Down**	**Up**
0	0.000122070313	0.000122070313	1.0
1	0.001586914063	0.001708984375	0.999877929688
2	0.009521484375	0.01123046875	0.998291015625
3	0.034912109375	0.046142578125	0.98876953125
4	0.087280273438	0.133422851563	0.953857421875
5	0.157104492188	0.29052734375	0.866577148438
6	0.20947265625	0.5	0.70947265625
7	0.20947265625	0.70947265625	0.5
8	0.157104492188	0.866577148438	0.29052734375
9	0.087280273438	0.953857421875	0.133422851563
10	0.034912109375	0.98876953125	0.046142578125
11	0.009521484375	0.998291015625	0.01123046875
12	0.001586914063	0.999877929688	0.001708984375
13	0.000122070313	1.0	
	0.000122070313		

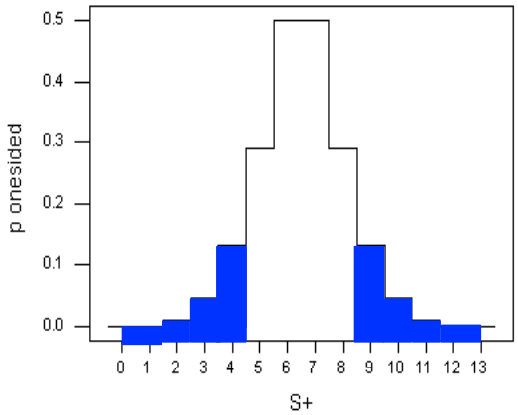


Figure B1. Two-sided critical region (p-values), shaded, for the sign test. n=13, S^+= 4 or 9.

Table B6 – Critical test statistic values for the signed-rank statistic W^+

(from McCornack, 1965)

© American Statistical Association. Used with permission.

The approximation given in chapter 6, used by most statistics software packages, can be used for $n > 15$ and $\alpha \geq 0.025$. For $\alpha < 0.025$, see exact tables in the McCornack paper or a textbook such as Hollander and Wolfe (1999), even for large sample sizes.

[reject H_0: at one-sided α when $W+ \leq$ table entry (small W)]

	α-level			
n	**.005**	**.010**	**.025**	**.050**
5				0
6			0	2
7		0	2	3
8	0	1	3	5
9	1	3	5	8
10	3	5	8	10
11	5	7	10	13
12	7	9	13	17
13	9	12	17	21
14	12	15	21	25
15	15	19	25	30
16	19	23	29	35
17	23	27	34	41
18	27	32	40	47
19	32	37	46	53
20	37	43	52	60

[reject H_0: at one-sided α when $W+ \geq$ table entry (large W)]

	α-level			
n	**.005**	**.010**	**.025**	**.050**
5				15
6			21	19
7		28	26	25
8	36	35	33	31
9	44	42	40	37
10	52	50	47	45
11	61	59	56	53
12	71	69	65	61
13	82	79	74	70
14	93	90	84	80
15	105	101	95	90
16	117	113	107	101
17	130	126	119	112
18	144	139	131	124
19	158	153	144	137
20	173	167	158	150

Table B7 – Critical test statistic values for the Friedman statistic Xf
(from Martin, Leblanc and Toan, 1993)

© The Canadian Journal of Statistics. Used with permission.

The chi-square approximation given in chapter 7 is used by most statistics software packages. For comparing 3 to 5 groups of data with sample sizes (blocks) n <10 in each group, an exact table should be used.

[reject H_0: at α when Xf $\geq$ table entry]

k = 3 groups

	α-level				
n	**.005**	**.010**	**.025**	**.050**	**.10**
3				6.00	6.00
4	8.00	8.00	8.00	6.50	6.00
5	10.00	8.40	7.60	6.40	5.20
6	10.33	9.00	8.33	7.00	5.33
7	10.29	8.86	7.71	7.14	5.43
8	9.75	9.00	7.75	6.25	5.25
9	10.67	8.67	8.00	6.22	5.56
10	10.40	9.60	7.80	6.20	5.00

k = 4 groups

	α-level				
n	**.005**	**.010**	**.025**	**.050**	**.10**
2				6.00	6.00
3	9.00	9.00	8.20	7.40	6.60
4	10.20	9.60	8.40	7.80	6.30
5	10.92	9.96	8.76	7.80	6.36
6	11.40	10.20	8.80	7.60	6.40
7	11.40	10.37	9.00	7.80	6.43
8	11.85	10.50	9.00	7.65	6.30
9	12.07	10.87	9.13	7.80	6.47
10	12.00	10.80	9.12	7.80	6.36

k = 5 groups

	α-level				
n	**.005**	**.010**	**.025**	**.050**	**.10**
2		8.00	8.00	7.60	7.20
3	10.67	10.13	9.60	8.53	7.47
4	12.00	11.20	9.80	8.80	7.60
5	12.48	11.68	10.24	8.96	7.68
6	13.07	11.87	10.40	9.07	7.73
7	13.26	12.11	10.51	9.14	7.77
8	13.50	12.30	10.60	9.30	7.80
9	13.69	12.44	10.67	9.24	7.73
10	13.84	12.48	10.72	9.28	7.76

Table B8 -- Quantiles (p-values) for Kendall's S statistic and tau correlation coefficient

For N>10 use the approximation given in section 8.2.2

One-sided p = Prob [S ≥ x] = Prob [S ≤ −x]

	N = Number of data pairs			
	4	5	8	9
x				
0	0.625	0.592	0.548	0.540
2	0.375	0.408	0.452	0.460
4	0.167	0.242	0.360	0.381
6	0.042	0.117	0.274	0.306
8		0.042	0.199	0.238
10		0.0083	0.138	0.179
12			0.089	0.130
14			0.054	0.090
16			0.031	0.060
18			0.0156	0.038
20			0.0071	0.022
22			0.0028	0.0124
24			0.0009	0.0063
26			0.0002	0.0029
28			<0.0001	0.0012
30				0.0004
32				0.0001
34				<0.0001
36				<0.0001

	N = Number of data pairs			
	3	6	7	10
x				
1	0.500	0.500	0.500	0.500
3	0.167	0.360	0.386	0.431
5		0.235	0.281	0.364
7		0.136	0.191	0.300
9		0.068	0.119	0.242
11		0.028	0.068	0.190
13		0.0083	0.035	0.146
15		0.0014	0.015	0.108
17			0.0054	0.078
19			0.0014	0.054
21			0.0002	0.036
23				0.023
25				0.0143
27				0.0083
29				0.0046
31				0.0023
33				0.0011
35				0.0005
37				0.0002
39				<0.0001
41				<0.0001
43				<0.0001
45				<0.0001

Table generated by D. Helsel

Appendix C

Data Sets

		Chapter cited
Data Set C1	Annual peak discharges for the Saddle River, NJ	2, 3
Data Set C2	Annual streamflows for the Conecuh River, AL	3, 6
Data Set C3	Daily streamflow for the Potomac River, Wash. D.C.	3
Data Set C4	Atrazine concentrations	6
Data Set C5	Subset of iron concentrations at low flow	7
Data Set C6	Complete set of iron concentrations	7
Data Set C7	Specific capacities of wells in Pennsylvania	7
Data Set C8	Corbicula on the Tennessee River	7
Data Set C9	TDS concentrations for the Cuyahoga River, Ohio	9
Data Set C10	Phosphorus transport, Illinois River at Marseilles	9
Data Set C11	Grain size and permeability of alluvial aquifers	9
Data Set C12	ROE and TDS data, Rappahannock R. near Fredericksburg, Virginia	10
Data Set C13	Streamflow data used for record extension	10
Data Set C14	Mean annual runoff and basin characteristics	11
Data Set C15	Urban total nitrogen loads	11
Data Set C16	Uranium and TDS in groundwaters	11
Data Set C17	Green River, Kentucky sediment transport data	12
Data Set C18	Maumee River, Ohio total P trends data	12
Data Set C19	Water levels, P-R-M system middle aquifer, NJ	12
Data Set C20	Factors affecting contamination from impoundments	15

Data sets are available in both ASCII and MS Excel formats. See the online location from which you obtained this book for the data files HhappC.dat and HhappC.xls .

Appendix D

Answers to Selected Exercises

Chapter 1

1.1 For the well yield data:

a) mean = 0.19
b) trimmed mean = 0.05
c) geometric mean = 0.04
d) median = 0.04
e) They differ because the data are skewed. The estimates which are more robust are similar, while the mean is larger.

1.2

a) standard deviation = 0.31
b) interquartile range = 0.36
c) MAD = 0.04
d) skew = 2.07. quartile skew = 0.83.

Because the data are asymmetric, the median difference is small, but the IQR and standard deviation are not.

1.3

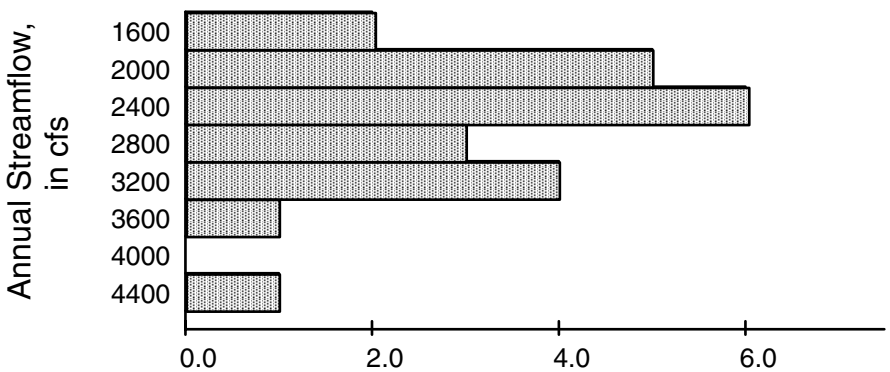
mean = 1.64	std. dev. = 2.85
median = 0.80	IQR = 0.61
geometric mean = 0.04	MAD = 0.25
skew = 3.09	quartile skew = –0.10

The largest observation is an outlier. Though the skew appears to be strongly positive, and the standard deviation large, this is due only to the effect of that one point. The majority of the data are not skewed, as shown by the more resistant quartile skew coefficient.

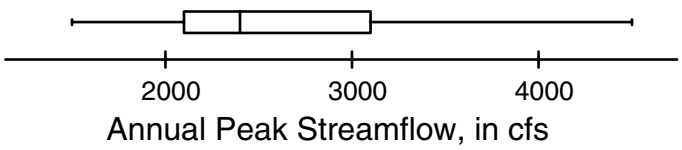
a) assuming the outlying observation is accurate, representing some high-nitrogen location which is important to have sampled, the mean must be used to compute the mass of nitrogen falling per square mile. It would probably be computed by weighting concentrations by the surface area represented by each environment. The median would under-represent this mass loading.
b) the median would be a better "typical" concentration, and the IQR a better "typical" variability, than the mean and standard deviation. This is due to the strong effect of the one unusual point on these traditional measures.

Chapter 2

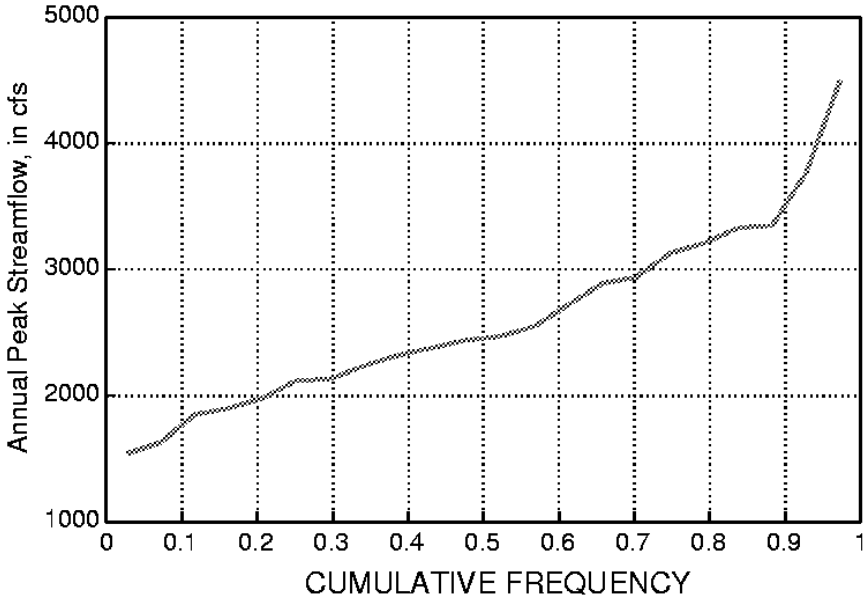
2.1 a)



b)

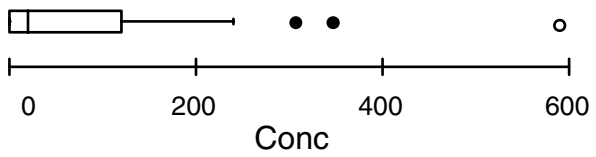


c)

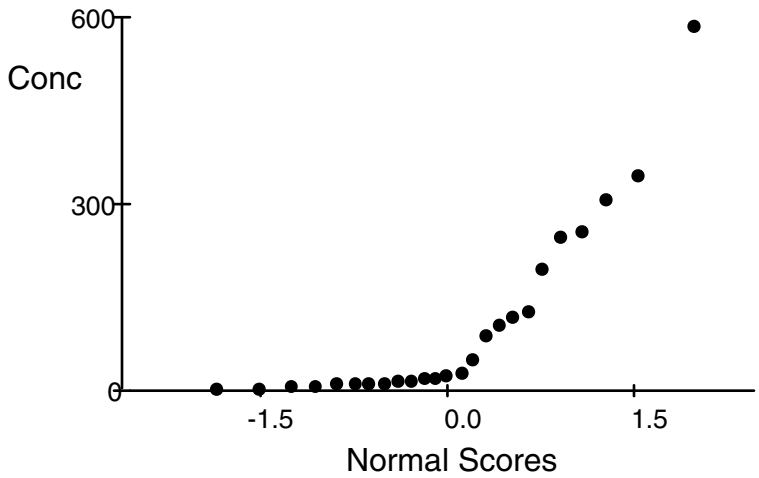


Either a cube root or logarithmic transform would increase symmetry.

2.2 a)

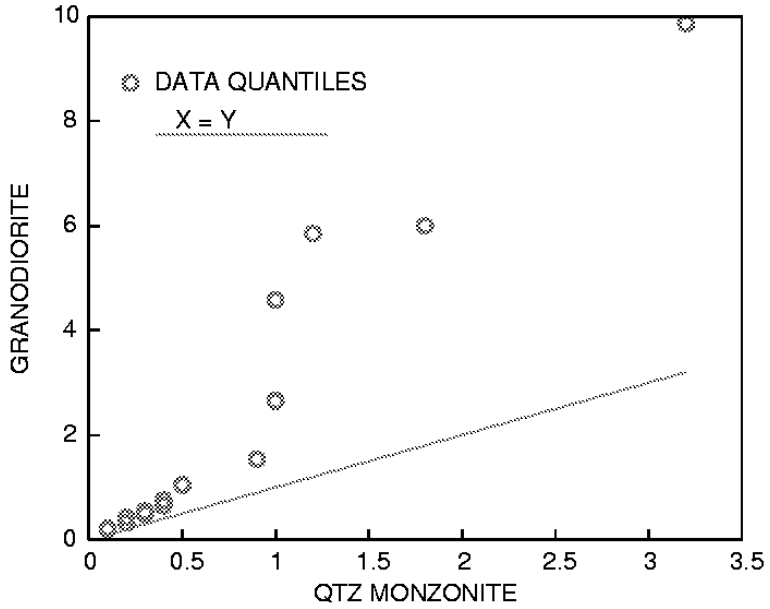


b)

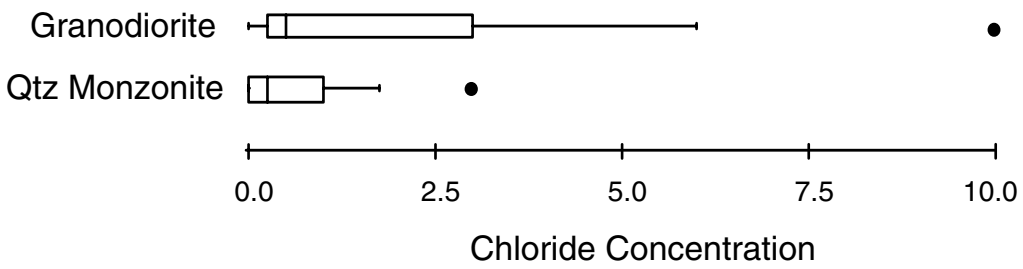


The data are strongly right-skewed. A log transformation makes these data more symmetric.

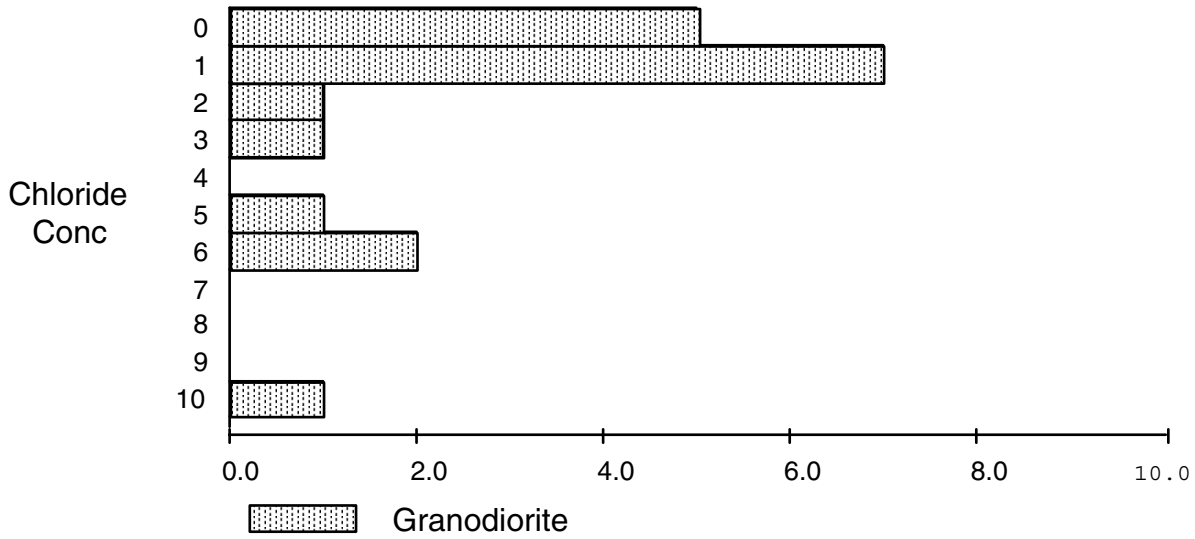
2.3 Q-Q plot.

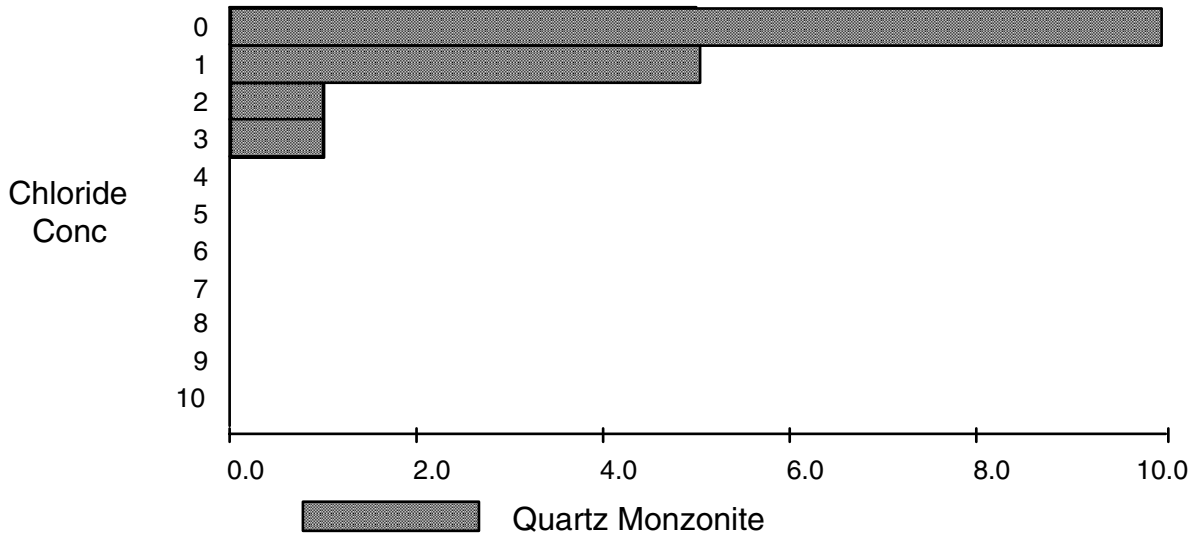


Boxplots:



Histograms:





The granodiorite shows higher chloride concentrations than does the quartz monzonite. This can be seen with any of the three graphs, but most easily with the boxplot. From the Q-Q plot, the relationship does not look linear.

2.4 There appears to be no effect of the waste treatment plant.

Chapter 3

3.1 nonparametric: $x' = 4$ ($\alpha/2 = .0154$). $R_l = 5$, $R_u = 14$.

$0.4 \le Cl_{0.5} \le 3.0$ at $\alpha = 0.031$.

This is as close to 0.05 as the table allows.

parametric: Using the natural logs of the data,

$$\exp\left(-0.045 - 2.11 \bullet \sqrt{1.63/18}\right) \le GM_x \le \exp\left(-0.045 + 2.11 \bullet \sqrt{1.63/18}\right).$$

$$0.51 \le GM_x \le 1.80.$$

Either of intervals is reasonable. The logs of the data still retain some skewness, so the nonparametric interval may be more realistic. The parametric interval might be preferred to obtain an alpha of 0.05. The choice would depend on whether the assumption of lognormality was believed.

3.2 symmetric: $0.706 - 2.12 \bullet \sqrt{0.639/17} \le \mu \le 0.706 + 2.12 \bullet \sqrt{0.639/17}$

$$0.30 \le \mu \le 1.12$$

Point estimates: mean = 0.705 (assuming normal distribution).

mean = $\exp(-0.849+0.5\bullet1.067)$

= 0.73 (assuming a lognormal distribution).

As the logs of the data are more symmetric than the data prior to transformation, the lognormal (2nd) estimate is preferred.

3.3 <u>Parametric 95% prediction interval:</u>

$0.19 - 2.20 \bullet \sqrt{0.0975 + (0.0975/12)}$ to $0.19 + 2.20 \bullet \sqrt{0.0975 + (0.0975/12)}$

or -0.53 to 0.91 gallons/min/foot. Includes 0.85, so same distribution.

<u>Nonparametric 95% prediction interval:</u>

$X_{[0.025\bullet13]}$ to $X_{[0.975\bullet13]}$ $X_{0.325}$ to $X_{12.675}$

The sample size is too small to produce such a wide (95%) nonparametric prediction interval. Therefore a parametric interval must be used. However, the negative lower end of the parametric prediction interval indicates that a symmetric interval is not appropriate. So an asymmetric interval resulting from taking logarithms should be used instead of the one above.

3.4 The data look relatively symmetric, so no logs taken.

mean:	683 ± 126, or	557->809	$\alpha = .05$.
median:	$R_l=6$, $R_u=15$	524->894	$\alpha = .041$.

3.5 The 90th percentile = 2445 cfs. A one-sided 95% confidence interval for the 90th percentile (an upper confidence limit to insure that the intake does not go dry) is found using the large-sample approximation of equation 3.17:

$$R_u = 365 \bullet 0.1 + z_{[0.95]} \bullet \sqrt{365 \bullet 0.1 \bullet (0.9)} + 0.5$$
$$= 36.5 + 1.645 \bullet 5.73 + 0.5 = 46.4$$

The 46th ranked point is the upper CI, or 2700 cfs.

Chapter 4

4.1 For the before-1969 data, PPCC r=0.986. For the after-1969 data, PPCC r=0.971. Critical values of r are 0.948 and 0.929, respectively. Therefore normality cannot be rejected for either period at $\alpha = 0.05$.

4.2 For the arsenic data, PPCC r=0.844. The critical r* from Appendix table B3 is r*=0.959. Therefore reject normality. For log-transforms of the data, PPCC r=0.973. Normality of the transformed data is not rejected.

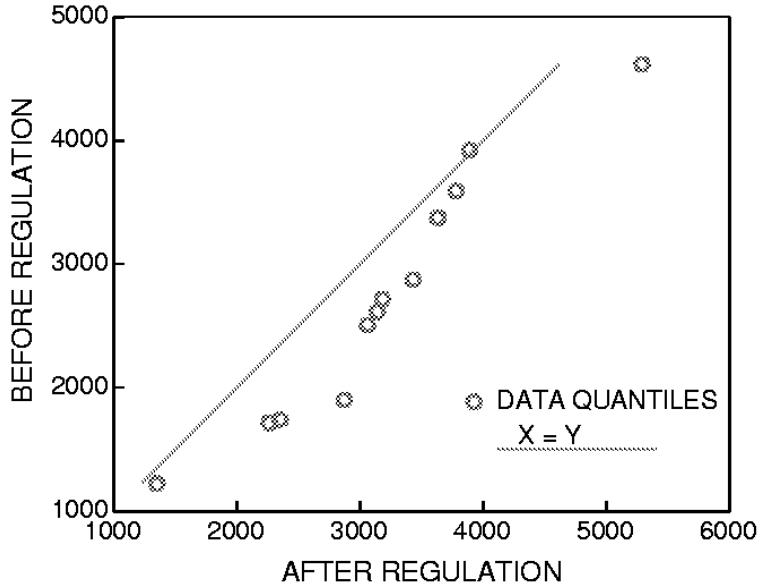
Chapter 5

5.1 The p-value remains the same.

5.2 Given that we wish to test for a <u>change</u> in concentration, but the direction of the change is not specified in the problem, this should be a two-sided test. If it had stated we were looking for an increase, or a decrease, the test would have been a one-sided test.

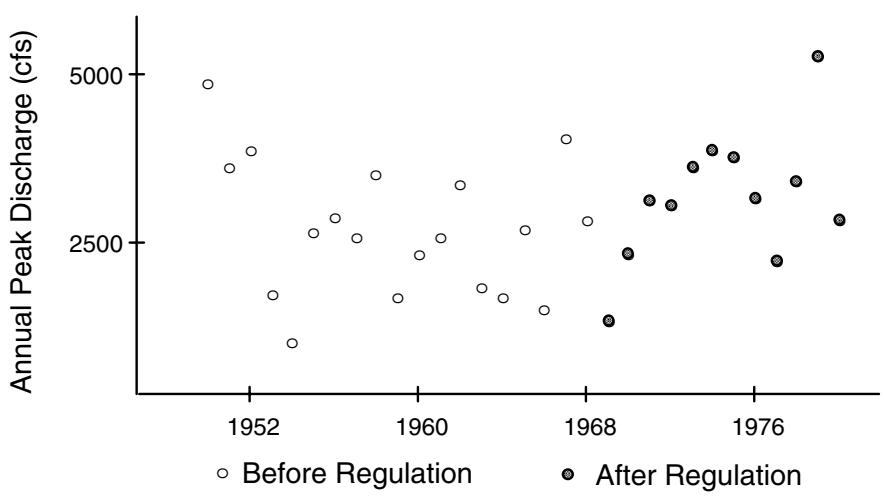
5.3 a. Quantiles are the 12 "after" data, and 12 quantiles computed from the 19 "before" data :

i	j	"after"	"before"
1	1.34	1350.00	1222.13
2	2.92	2260.00	1715.25
3	4.49	2350.00	1739.84
4	6.06	2870.00	1900.82
5	7.64	3060.00	2506.23
6	9.21	3140.00	2614.92
7	10.79	3180.00	2717.21
8	12.36	3430.00	2873.61
9	13.93	3630.00	3375.24
10	15.51	3780.00	3591.15
11	17.08	3890.00	3922.29
12	18.66	5290.00	4617.37



The relationship appears additive. The Hodges-Lehmann estimate (median of all possible after–before differences) = 480 cfs.

b. After regulation, the reservoir appears to be filling. Any test for change in flow should omit data during the transition period of 1969-1970. Plots of time series are always a good idea. They increase the investigator's understanding of the data. Low flows after regulation are not as low as those before. This produces the pattern seen in the Q-Q plot of the low quantiles being lower after regulation. while the upper quantiles appear the same, as shown by the drift closer to the x=y line for the higher values.



c. With 1969 and 70 included, W_{rs} = 273.5 p=0.22. The after flows are not significantly different. With 1969 and 70 excluded, W_{rs} = 243.5 p=0.06. The after flows are close to being significantly different -- more data after regulation is needed.

5.4 Exact test

X	Y	R(Y)	R(X)
1			1
	1.5	2	
2			3
	2.5	4	
3			5
	3.5	6	
4			7
	4.5	8	
	5.5	9	
	7.0	10	
	10.0	11	
	20.0	12	
	40.0	13	
	100.0	14	

$n = 4 \quad m = 10 \qquad W_{rs} = \Sigma R_x = 16$

From table B4, Prob($W_{rs} \leq 16$) = .027. The two-sided exact p-value = 0.054

Large-sample approximation

The mean is $\mu_W = \dfrac{n \bullet (N+1)}{2} = \dfrac{4 \bullet 15}{2} = 30$

The standard deviation is given by $\sigma_W = \sqrt{\dfrac{n \bullet m \bullet (N+1)}{12}} = 7.0711$

$$Z_{rs} = \frac{16 - \mu_W + 1/2}{\sigma_W} = -1.909$$

Using linear interpolation between −1.9110 and −1.8957 in a table of the standard normal distribution gives the one-tail probability of 0.028. So the two-sided approximate p-value is 0.056.

t-test on the ranks

Replacing variable values by ranks gives

$\bar{x} = 4 \qquad S_x = 2.582 \qquad S^2_x = 6.667 \qquad n = 4$

$\bar{y} = 8.9 \qquad S_y = 3.928 \qquad S^2_y = 15.429 \qquad m = 10$

The pooled variance is :

$$S^2 = \frac{3S^2_x + 9S^2_y}{12} = 13.2386$$

$S = 3.639$

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{(1/n + 1/m)}} = -2.27610$$

Linear interpolation for a student's t with 12 degrees of freedom gives

$$.975 + \frac{(2.27610 - 2.1788)}{(2.6810 - 2.1788)} \bullet .015 = .97791 \qquad 1.0 - .97791 = .022$$

The two-sided rank transform p-value is .044.

Summary	
Approach	**p-value**
Rank-Sum Exact	0.054
Rank-Sum Approx.	0.056
t test on ranks	0.044

To compute $\hat{\Delta}$, the (n•m)=40 differences ($X_i - Y_j = D_j$) are:

(Y_1)	0.5	1.5	2.5	**3.5**	4.5	6	9	19	39	99
(Y_2)	−0.5	0.5	1.5	2.5	3.5	5	8	18	38	98
(Y_3)	−1.5	−0.5	0.5	1.5	2.5	**4**	7	17	37	97
(Y_4)	−2.5	−1.5	−0.5	0.5	1.5	3	6	16	36	96

$\hat{D}$ = median of 40 D_j's $(D_{\text{rank } 20} + D_{\text{rank } 21})/2 = 3.75$

5.5 Yields with fracturing: $r_{crit} = .932$, accept normality

Yields without: $r_{crit} = .928$, reject normality

Because one of the groups is non-normal, the rank-sum test is performed. $W_{rs} = \Sigma R_{without} = 121.5$. The one-sided p-value from the large-sample approximation p= 0.032. Reject equality. The yields from fractured rocks are higher.

5.6 The test statistic changes very little ($W_{rs} = 123$), indicating that most information contained in the data below detection limit is extracted using ranks. Results are the same (one-sided p-value = 0.039. Reject equality). A t-test could not be used without improperly substituting some contrived values for less-thans which might alter the conclusions.

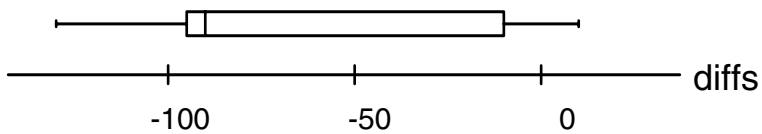
Chapter 6

6.1 The sign test is computed on all data after 683 cfs is subtracted. $S^+ = 11$. From table B5, reject if $S^+ \geq 14$ (one-sided test). So do not reject. $p > 0.25$.

6.2 c is not a matched pair.

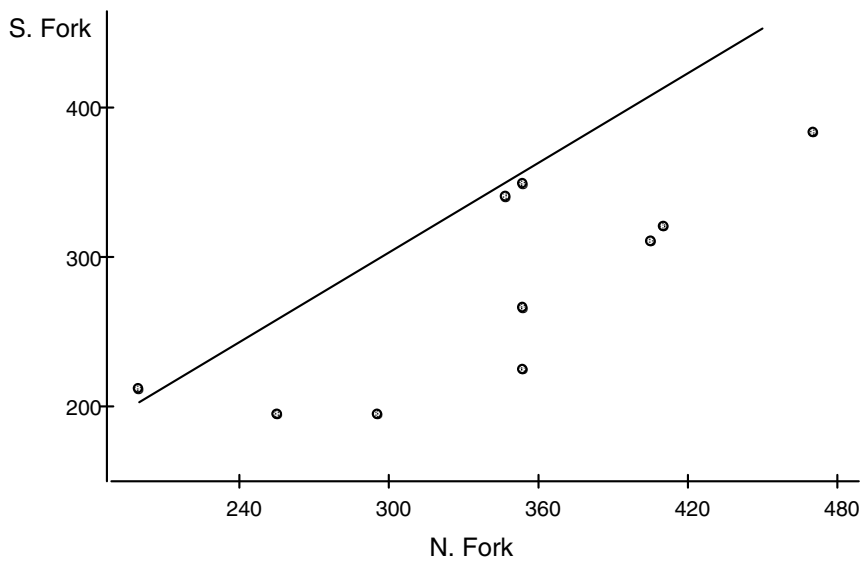
6.3 a. H_0: μ (South Fork) – μ (North Fork) = 0.
H_1: μ (South Fork) – μ (North Fork) $\neq$ 0.

b. A boxplot of the differences shows no outliers, but the median is low. Conductance data are usually not skewed, and the PPCC r=0.941, with normality not rejected. So a t-test on the differences is computed (parametric).



c. t = –4.24 p = 0.002 Reject H_0.

d. Along with the boxplot above, a scatterplot shows that the South Fork is higher only once:



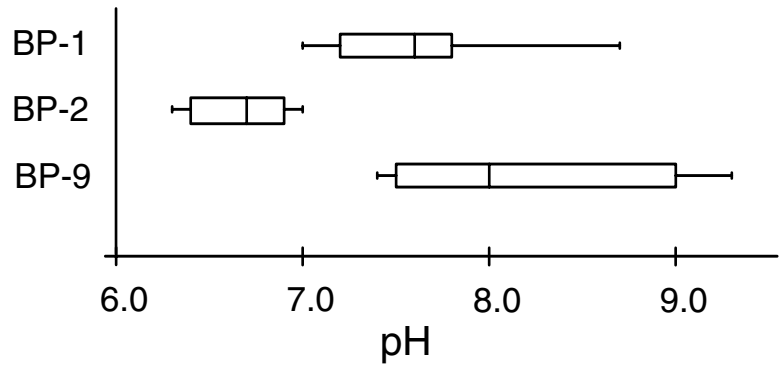
e. The mean difference is –64.7.

6.4 Because of the data below the reporting limit, the sign test is performed on the differences Sept–June. The one-sided p-value = 0.002. Sept atrazine concentrations are significantly larger than June concs before application.

6.5 For the t-test, t=1.07 with a one-sided p-value of 0.15. The t-test cannot reject equality of means because one large outlier in the data produces violations of the assumptions of normality and equal variance.

Chapter 7

7.1 As a log-transformed variable, pH often closely follows a normal distribution. See the following boxplots:



pH for three piezometer groups (from Robertson et al., 1984)

The PPCC for the three groups (0.955 for BP-1, 0.971 for BP-2, and 0.946 for BP-9) cannot disprove this assumption. Therefore ANOVA will be used to test the similarity of the three groups.

Anova Table:

Source	df	SS	MS	F	p-value
Piez Gp	2	7.07	3.54	9.57	0.002
Error	15	5.54	0. 37		
Total	17	12.61			

The groups are declared different. Statistics for each are:

GP	N	Mean	Std. Dev.	Pooled Std. Dev = 0.608
BP-1	6	7.65	0.596	
BP-2	6	6.68	0.279	
BP-9	6	8.20	0.822	

A Tukey's test on the data is then computed to determine which groups are different. The least significant range for Tukey's test is

$$\text{LSR} = q_{(0.95,\,2,\,15)} \bullet \sqrt{0.37/6} = 3.01 \bullet 0.248$$
$$= 0.75$$

Any group mean pH which differs by more than 0.75 is significantly different by the Tukey's multiple comparison test. Therefore two piezometer groups are contaminated, significantly higher than the uncontaminated BP-2 group:

$$\text{BP-9} \cong \text{BP-1} > \text{BP-2}$$

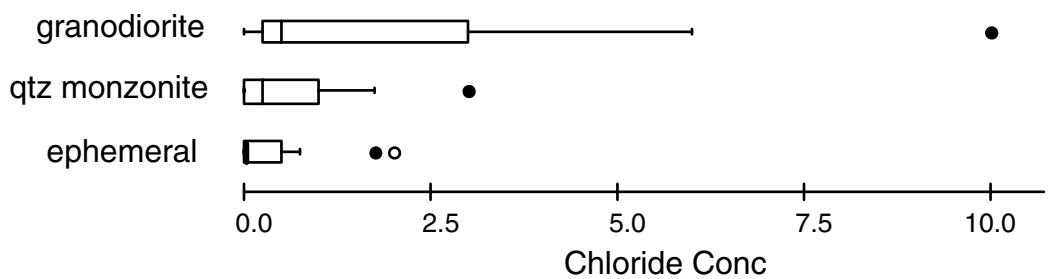
Since the sample sizes are small (n=6 for each group) one might prefer a Kruskal-Wallis test to protect against any hidden non-normality:

GP	N	MEDIAN	$\bar{R}_j$
BP-1	6	7.60	11.3
BP-2	6	6.75	3.6
BP-9	6	8.00	13.6

Overall Median = 9.5

K = 11.59 $\chi^2_{0.95,(2)} = 5.99$. Reject H_0, with p = 0.003.
ANOVA and Kruskal-Wallis tests give identical results.

7.2 Boxplots of the data indicate skewness. Therefore the Kruskal-Wallis test is computed:
K = 7.24 Corrected for ties, K = 7.31. p = 0.027
Reject that all groups have the same median chloride concentration.



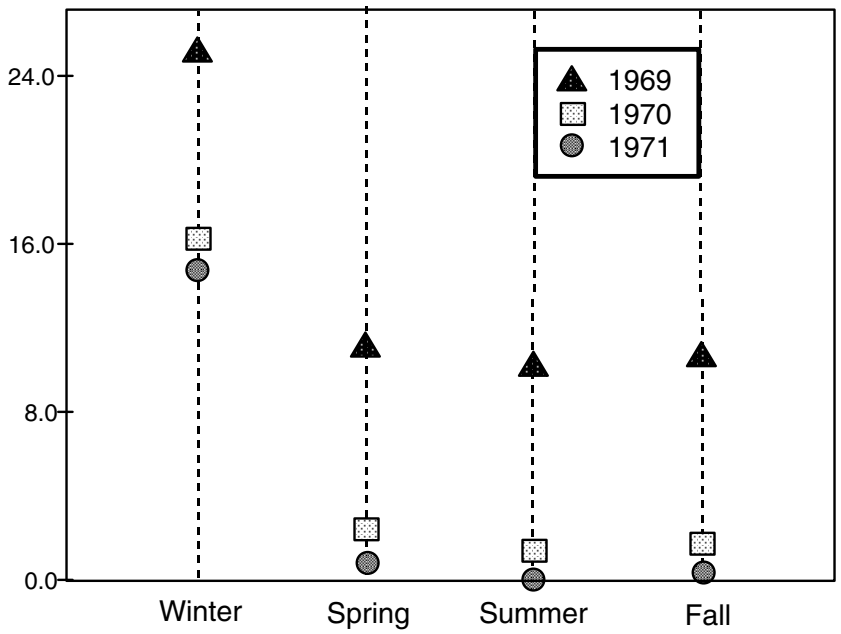
The medians are ranked as granodiorite > qtz monzonite > ephemeral. Individual K-W tests are computed for adjacent pairs at $\alpha = 0.05$:
granodiorite ≅ qtz monzonite (p = 0.086)
qtz monzonite ≅ ephemeral (p = 0.27). So:

granodiorite qtz monzonite ephemeral

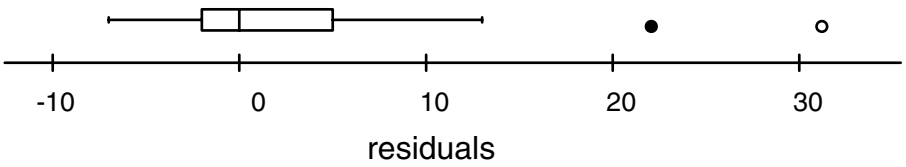
7.3 Median polish for the data of strata 1:

	Winter	Spring	Summer	Fall	Year median
1969	25.25	11.25	10.25	10.75	8.75
1970	16.5	2.5	1.5	2	0.00
1971	15	1	0	0.5	−1.50
Season median	14.25	0.25	−0.75	−0.25	**2.25**

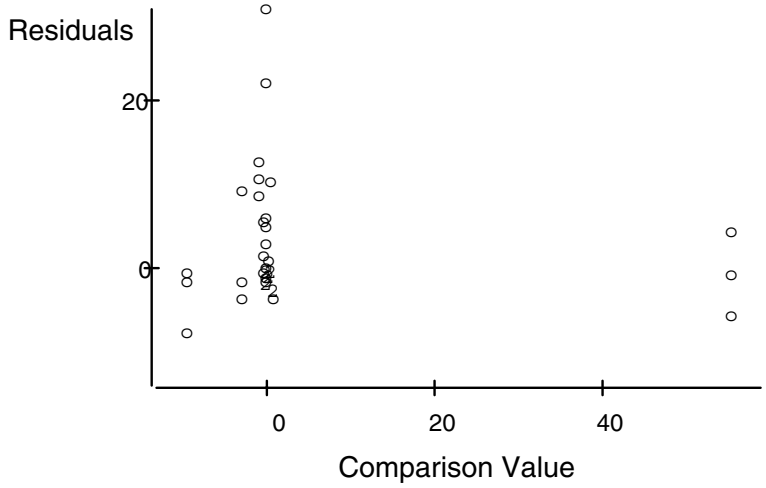
Corbicula densities were 14 units higher in winter than in other seasons, and 9 to 10 units higher in 1969 than 1970 or 1971. Those effects dominated all others. This is shown by a plot of the two-way polished medians:



The residuals are skewed, as shown in a boxplot:



However, a residuals plot of cell residuals versus the comparison value shows outliers, but an overall slope of zero, stating that no power transformation will improve the situation very much.



7.4 Due to the outliers noted above, ranks of the Corbicula data were used to test the effects of season and year on the number of organisms.

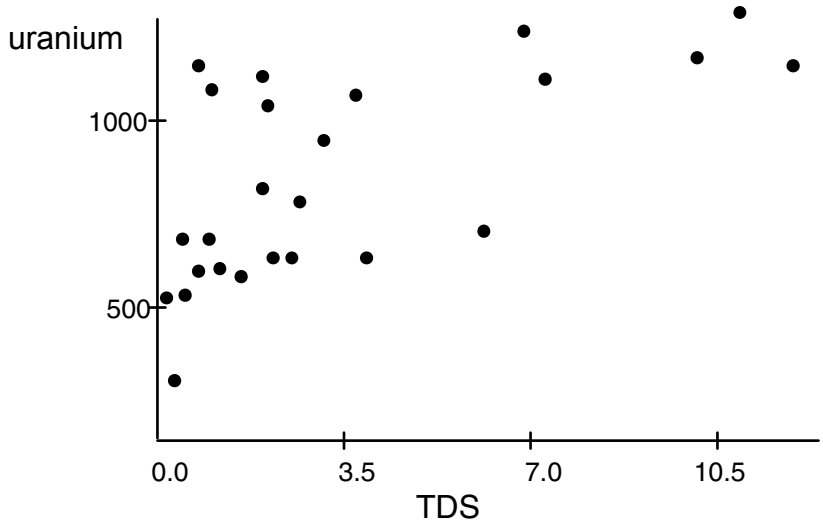
Source	df	SS	MS	F	p-value
Year	2	1064.7	532.33	13.78	0.000
Season	3	1300.7	433.57	11.23	0.000
Year*Season	6	560.8	93.46	2.42	0.057
Error	24	926.8	38.62		
Total	35	3853.0			

A two-way ANOVA on the ranks indicates that both season and year are significant influences on the density of Corbicula, and that there is no interaction. This is illustrated well by the plot of polished medians above.

7.5 Not answered.

Chapter 8

8.1 The plot of uranium versus total dissolved solids looks like it could be nonlinear near the 0 TDS boundary. So Spearman's rho was computed, and
rho = 0.72 with t_r = 4.75 and p<0.001.



8.2 Pearson's r = 0.637 with t_r = 3.79 and p<0.001. Kendall's tau = 0.53 with p<0.001. The suggestion of nonlinearity would favor either rho or tau, though the nonlinearity is not serious in this case.

8.3 Not answered.

Chapter 9

9.1 A residuals plot for the untransformed variables shows strong curvature. A log-log regression gives an acceptable plot, with one outlier not influencing the line:

log(Yield) = 6.74 + 1.39•log(Grain Size) t= 8.14 $p < 0.001$

9.2 The overall mean yield will be the average of estimates of mean yield for the four wells from the regression equation. Applying the 1/2 s^2 correction factor to obtain the mean yield rather than the median, the estimated mean yields are:
46.104 120.830 316.669 556.380 with overall mean = 260 gal/day/ft^2.

9.3 Here are some possible transformations, including the log. Can logQ be improved on?

explanatory variable	R^2
Q	40.8%
$Q^{0.5}$	51.1%
log Q	57.3%
$Q^{-0.25}$	57.4%
$Q^{-0.5}$	55.4%
Q^{-1}	47.9%
1/(1+0.00001Q)	41.8%
1/(1+0.0001Q)	47.6%
1/(1+0.001Q)	58.5%
1/(1+0.01Q)	52.4%
1/(1+0.1Q)	48.5%

There are perhaps two other good candidate explanatory variables on this list, $Q^{-0.25}$, and 1/(1+0.001Q). Neither improve significantly over logQ, based on R^2 or on residuals plots. A residuals plot and probability plot of residuals for the hyperbolic transformation having b=0.001 are below.

When b=0.00001 or smaller, the model is virtually identical to the linear model TDS = b_0+b_1Q [a power transformation with $\theta = 1$]. When b=0.1 or larger, the model is virtually identical to the inverse model TDS = $b_0+b_1(1/Q)$ [a power transformation with $\theta = -1$]. Values of b in between these provide functions similar to moving down the ladder of powers from $\theta = 1$ to $\theta = -1$. The advantage of using the hyperbolic function is its interpretability as a mixing of ground and surface waters (Johnson et al., 1969).

9.4 If the objective is to predict LOAD, then that (or its transform) should be the dependent variable. The regression statistics (especially PRESS) will then tell how well the predictions will do. If ln(C) is used as the dependent variable, the standard error s = 0.3394, exactly

the same as in the equation for ln(LOAD), but R^2=17.3 % rather than 67.9 % for ln(LOAD). The t statistic on β_1 is −4.43, also significant but not as much as when y = ln(LOAD). In other words, the error of the ln(C) values is exactly the same magnitude as the errors of ln(LOAD). The percent variation explained drops from 67.9 % to 17.3 %, the difference being the strong effect of Q on variation in LOAD. Note the changes in regression coefficients. The previous model was ln(LOAD) = 0.789 + 0.761 ln(Q). This one is ln(C) = −.194 − .239 ln(Q). The intercept decreased by an amount equal to ln(2.7) (the log of the unit conversion coefficient) and the slope decreased by exactly 1 because Q is removed from both sides. The standard errors of the coefficients are both unchanged.

If LOAD were computed by using the regression for ln(C) and then multiplying that result by 2.7 Q, exactly the same estimates would result as when using the equation for ln(LOAD). This is true regardless of which estimation method is employed (median, MLE, or Smearing), and will always be true for log-log regression estimation. The moral of the story is: if your boss thinks that you shouldn't use ln(LOAD) as the dependent variable and you can't convince him or her otherwise, go ahead and predict ln(C), and from that ln(LOAD), and you will still get the results you got doing it the simple way.

Chapter 10

10.1

X	Y	Slopes				
1	10	30	10	15	13	9.2
2	40	−10	7.5	7.33	4	
3	30	25	16	8.67		
4	55	7	0.5			
5	62	−6				
6	56					

Ranked slopes: −10, −6, 0.5, 4, 7, 7.33, 7.55, 8.67, 9.2, 10, 13, 15, 16, 25, 30

a) Median slope = 8.67 = Theil slope estimator $\hat{b}_1$
Median X = 3.5
Median Y = 47.5
S = P − M = 13 − 2 = 11

b) $\tau = \dfrac{S}{n(n-1)/2} = \dfrac{11}{6 \bullet 5/2} = 0.73$

c) intercept $\hat{b}_0 = Y_{med} - \hat{b}_1 \bullet X_{med} = 47.5 - 8.67 \bullet 3.5 = 17.17$
Y = 17.17 + 8.67• X is the Kendall-Theil equation
(Y = 10.07 + 9.17• X is the OLS equation for the same data)

d) from table B8, for S=11 and n=6, two-sided p-value = 2•0.028 = 0.056.

10.2

X	Y	Slopes				
1	10	30	10	15	47.5	9.2
2	40	−10	7.5	53.33	4	
3	30	25	85	8.67		
4	55	145	0.5			
5	200	−144				
6	56					

Ranked slopes −144, −10, 0.5, 4, 7.5, 8.67, 9.2, 10, 15, 25, 30, 47.5, 53.33, 85, 145

a) Median slope = 10 = Theil slope estimator $\hat{b}_1$
Median X = 3.5
Median Y = 47.5

b) S and τ are unchanged

c) $\hat{b}_0 = 47.5 - 10 \bullet 3.5 = 12.5$
Y = 12.5 + 10 • X the Kendall-Theil equation is similar to ex. 10.1.
(Y = −8.33 + 21 • X the OLS slope has changed a lot from ex. 10.1)

d) the p-value is unchanged.

e) for a 95% confidence interval, the closest entry in table B8 to $\alpha/2$=0.025 is p=0.028 for X_u=11. From eqs. 10.3 and 10.4,

$$R_u = \frac{(15 + 11)}{2} = 13 \text{ for N=15 and } X_u=11.$$

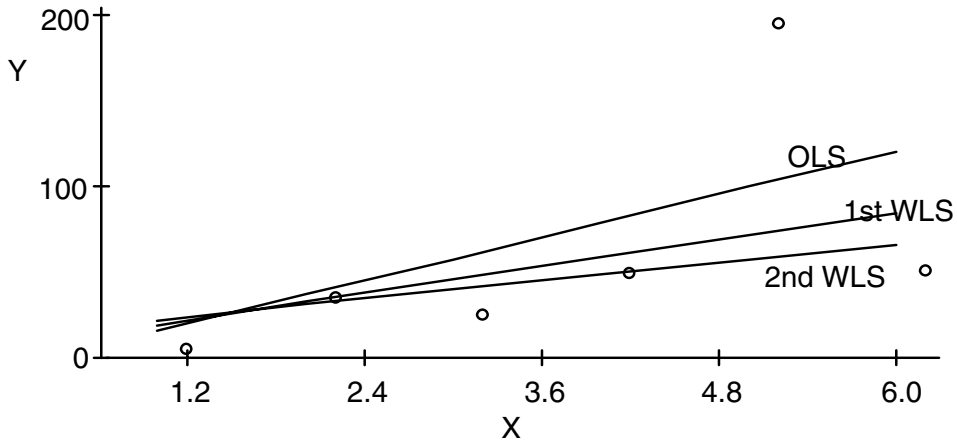
The rank R_l of the pairwise slope corresponding to the lower confidence limit is

$$R_l = \frac{(15 - 11)}{2} + 1 = 3.$$

So an α = 0.054 confidence limit for $\hat{\beta}_1$ is the interval between the 3rd and 13th ranked pairwise slope (the 3rd slope in from either end), or

$$0.5 \le \hat{\beta}_1 \le 53.3 .$$

10.3 The unweighted OLS regression equation is
Y = − 8.3 + 21.0•X t = 1.41 p = 0.23



The residuals are then divided by 6•(MAD), where the MAD is the median of the absolute values of the residuals. Bisquare weights are computed for each data point:

pt #	1	2	3	4	5	6
weight	0.999	0.996	0.935	0.954	0.179	0.631

A first weighted least squares is then computed:

$Y = 3.1 + 13.1\ X$ $\quad t = 1.49 \quad p = 0.21$

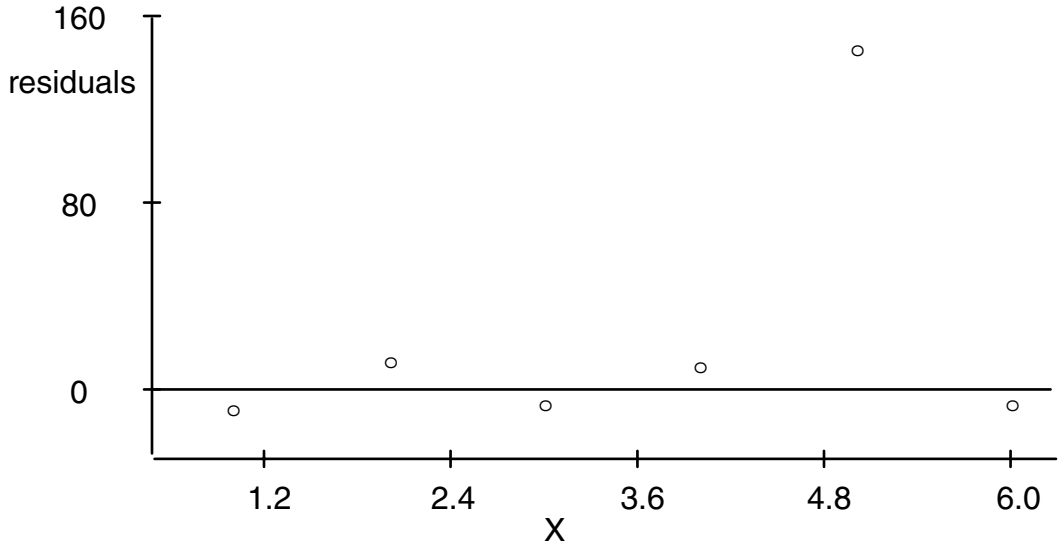
Bisquare weights are again computed for each data point, using residuals from the first WLS:

pt #	1	2	3	4	5	6
weight	0.984	0.952	0.938	1.000	0.000	0.746

A second WLS is then computed:

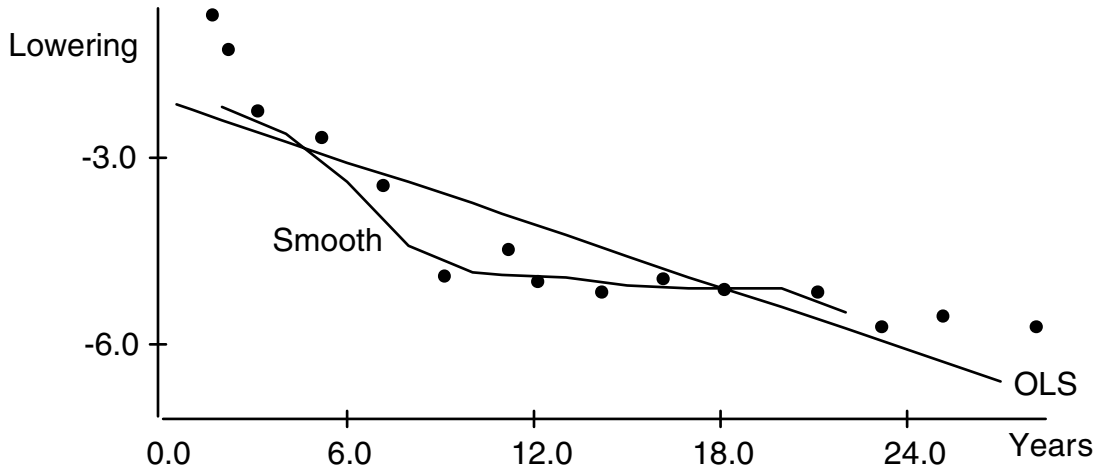
$Y = 10.4 + 8.80\ X$ $\quad t = 2.73 \quad p = 0.07$

Though the slope has diminished from the OLS line, the significance has greatly increased due to the lower weight of the outlier. Note the similarity between this WLS and the Kendall's robust line. A residuals plot shows that the WLS line fits most of the data much better than with OLS. The outlier's influence on the slope has diminished, and its residual remains large.



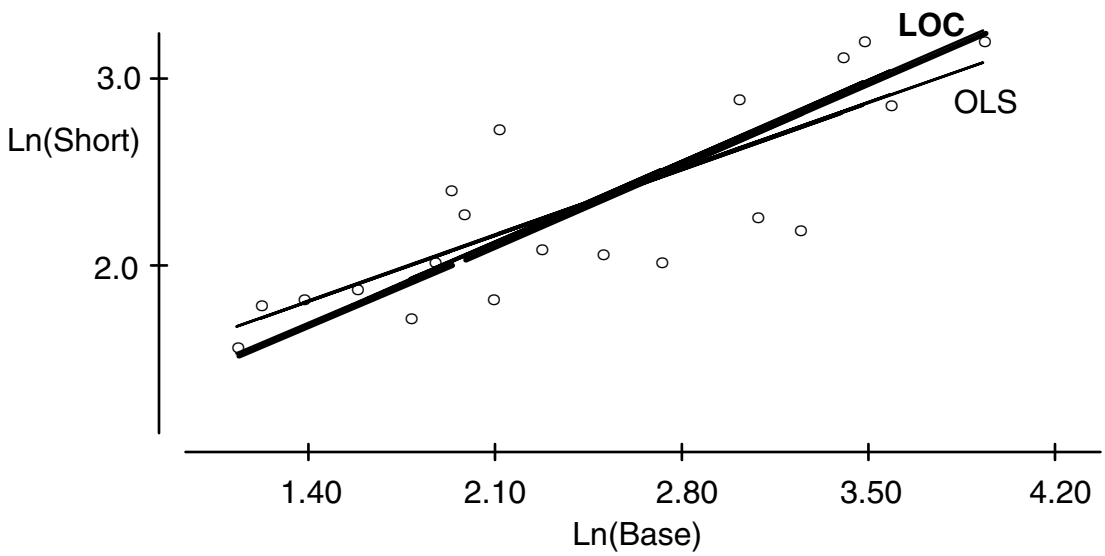
Some object: "Isn't this WLS line the same as throwing away the outlier -- it has a weight of zero?" The difference is that the outlier was determined to be downweighted to zero by the data itself, not an arbitrary decision by the data analyst. Weighted least squares also allows outliers to have partial weights, not simply a zero or one weight as with discarding the outlier. So WLS is far less arbitrary and far more consistent in its assignment of weights to all data points than is throwing away outliers.

10.4 Lowering = –2.07 – 0.167 •Years $r^2 = 0.76$



OLS does not follow the data as well as the smooth because the data are nonlinear.

10.5 Plotting the 20 years of joint data shows that curvature and heteroscedasticity exist, and transformation is required before regression. Thus the natural logs of both are taken. A linear relation results, as shown in the following plot.



Regression between the 20 year joint record at the two stations is:

$$\text{Ln(Short)} = 1.095 + 0.507 \bullet \text{Ln(Base)} \quad t = 6.00 \qquad p < 0.001 \qquad R^2 = 0.67$$

Using this equation and the 30 additional years of record at Base, 30 years of simulated flows at Short are generated. Now the LOC is used to generate estimates of the "Short" 30-year record. Summary statistics for the 20 years of joint Ln(Base) and Ln(Short) records are as follows:

	n	Mean	Stdev	Median	P25	P75
Ln(Short)	20	2.319	0.524	2.160	1.862	2.850
Ln(Base)	20	2.414	0.844	2.190	1.802	3.200

From equation 10.10,

$$Y_i = \overline{Y} + \text{sign}[r] \bullet \frac{s_y}{s_x} \bullet (X_i - \overline{X}), \quad \text{or}$$

$$\text{Ln(Short)} = 2.319 + (.524/.844) \bullet (\text{Ln(Base)} - 2.414)$$

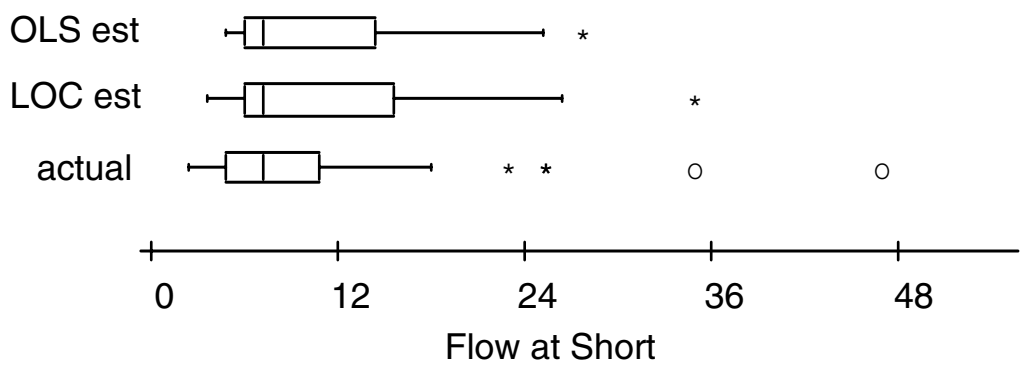
$$= 0.820 + .621 \bullet \text{Ln(Base)}$$

Note how the slope and intercept for LOC differ from the regression coefficients. Summary statistics for the estimated flows at "Short" by the two methods are compared to the true 30-year record from Appendix C13 in the following table.

	n	Mean	Stdev	Median	P25	P75
OLS est.	30	2.2087	0.4975	2.0228	1.7731	2.6249
LOC est.	30	2.184	0.609	1.956	1.651	2.694
true values	30	2.079	0.613	1.930	1.630	2.290

The standard deviation for the regression estimate is too small, as expected.

Boxplots are presented below for three groups: the 30-year estimates using regression and LOC combined with the 20-year record at Short, and the actual 50-year record. LOC comes closer to correctly estimating the lowest and highest flows. The regression estimates are too low for high flows, and too high for low flows. They "regress" toward the mean more than the actual data because the standard deviation of the estimates is too small, as $R^2 < 1$.



10.6 Not answered.

Chapter 11

11.1 The full multiple regression model contains strong multi-collinearity. The VIFs among the four percentage variables are huge:

LOGTN = – 1.3 + 0.596 LOGDA + 0.346 LOGIMP + 0.0314 MMJTEMP – 0.0494 MSRAIN + 0.040 PRES + 0.035 PNON + 0.037 PCOMM + 0.024 PIND

n = 42 s = 0.61 R^2 = 0.59

Parameter	Estimate	Std.Err(β)	t-ratio	p	VIF
Intercept β_0	–1.28	24.60	–0.05	0.959	
Slopes β_k					
LOGDA	0.596	0.121	4.94	0.000	1.8
LOGIMP	0.346	0.228	1.52	0.138	3.8
MMJTEMP	0.031	0.019	1.65	0.107	10.1
MSRAIN	–0.049	0.021	–2.32	0.026	9.1
PRES	0.040	0.245	0.16	0.873	9227.2
PNON	0.035	0.246	0.14	0.888	3062.2
PCOMM	0.037	0.245	0.15	0.882	8311.4
PIND	0.024	0.246	0.10	0.922	2026.2

Table 11.9 Regression statistics and VIF's for Exercise 11.1

To determine why the multi-collinearity is so strong, the correlation matrix is computed.

	LOGTN	LOG LOGDA	MMJ IMP	MS TEMP	RAIN	PRES	PNON	PCOMM
LOGDA	0.565							
LOGIMP	0.058	-0.382						
MMJTEMP	-0.205	-0.188	0.094					
MSRAIN	-0.259	-0.083	0.018	0.915				
PRES	0.294	0.210	-0.246	0.040	0.003			
PNON	-0.042	0.319	-0.639	0.066	0.065	-0.321		
PCOMM	-0.218	-0.441	0.589	-0.027	0.039	-0.747	-0.206	
PIND	-0.131	0.060	0.114	-0.111	-0.164	-0.226	-0.124	-0.180

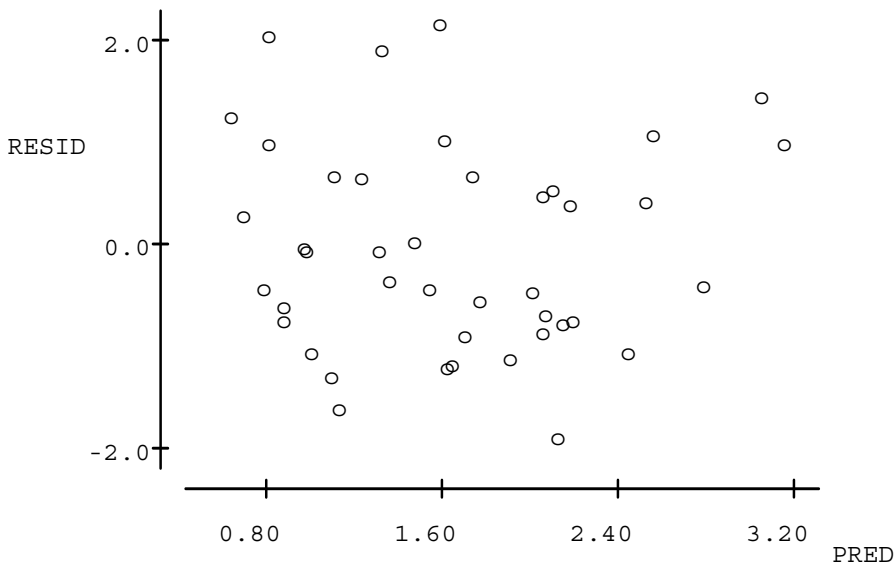
Surprisingly, the percentage terms do not have large pair-wise correlation coefficients. Instead, they are strongly related in that the four of them add to 100%, except for rounding error. This is why the VIFs are so large. Therefore at least one of them should be dropped. The variable with the smallest partial F (PIND) could be chosen. This brings the VIF down from over 9000 to 10, still large. In order to save much time the Cp and PRESS statistics can be computed for all possible models. The results below show that the best 5-variable model, containing LOGDA, LOGIMP, MMJTEMP,

MSRAIN, and PIND, is the best in terms of prediction errors (PRESS) and model bias/standard error (Cp). VIFs are below 10 ($R^2 < 0.9$) and so are acceptable.

X: Variable is in the model

# of Vars	R-sq	PRESS	C-p	Max VIF	LOGDA	LOGIMP	MJTEMP	MSRANN	PRESS	PNOISE	PCOMM	PIND
1	32.0	22.9	17.0	----	X							
1	8.6	30.0	35.8	----					X			
2	40.7	21.2	11.9	1.2	X	X						
2	37.5	22.2	14.5	1.1	X					X		
3	46.4	20.2	9.3	1.2	X	X			X			
3	45.2	19.8	10.3	1.2	X	X						X
4	51.2	18.9	7.4	1.2	X	X		X				X
4	50.9	19.8	7.7	1.2	X	X		X	X			
* 5	57.0	17.8	4.7	6.8	X	X	X	X				X
5	56.2	18.0	5.4	7.5	X		X	X	X		X	
6	59.1	18.2	5.1	7.1	X	X	X	X	X			X
6	58.6	18.9	5.4	7.1	X	X	X	X		X		X
7	59.1	19.4	7.0	10.0	X	X	X	X	X	X	X	
7	59.1	19.4	7.0	13.2	X	X	X	X	X		X	X
8	59.1	20.1	9.0	9227	X	X	X	X	X	X	X	X

A residuals plot from the minimum PRESS and Cp model shows no hint of curvature or increasing variance. Therefore this model is preferred.



Residuals plot for the regression of Exercise 11.1

11.2 First compare those models which have equal numbers of parameters and eliminate the ones with higher SSE.

Compare 4 to 7 , eliminate 4
Compare 3 to 6 , eliminate 3
Compare 2 to 5 , eliminate 2
Compare 6 to 8 , eliminate 6

Now, for the remaining models (1, 5, 7, 8, 9. 10) perform F tests between pairs of nested models. The order in which to proceed is arbitrary.

Compare	F	df_{num}	df_{denom}	F_{crit}	conclusion
Models 1 to 5	11.18	1	123	3.9	reject H_0,eliminate model 1
Models 5 to 7	0.28	2	121	3.1	do not reject eliminate model 7
Models 5 to 8	1.39	1	122	3.9	do not reject eliminate model 8
Models 5 to 9	0.77	3	120	2.68	do not reject eliminate model 9
Models 5 to 10	0.88	5	118	2.29	do not reject eliminate model 10

Model 5 is the preferred model.

Another possible approach is to use either PRESS or Mallows Cp.

Model	p	s^2	Cp
1	3	0.5636	14.29
2	4	0.5350	8.37
3	5	0.5343	9.17
4	6	0.5359	10.51
5	4	0.5183	4.41
6	5	0.5207	5.98
7	6	0.5245	7.84
8	5	0.5166	5.00
9	7	0.5212	8.06
10	9	0.5208	9.95

The results are interpreted as: the transport curve is quadratic with a shift in intercept for the winter months. Only two seasons (not three) can be distinguished. The slope of the curve does not change with season.

11.3 Not answered.

11.4 Not answered.

Chapter 12

12.1 Regression

load = 25,250 − 12.6 year r^2 = 10.6%

(t) (1.53) (−1.50) two-sided p value = 0.134

Multiple regression

load = 28,152 − 14.4 year + 0.696 q r^2 = 88.3%

(t) (4.69) (−4.60) (10.91) two-sided p value ≅ 0.0001

Mann-Kendall

load = 11,800 − 5.8 year two-sided p value = 0.415

Mann-Kendall on Residuals

Regression model is load = −110 + 0.681 q r^2 = 74.5%

(t) (−1.24) (7.44)

Kendall fit: residual = 28,250 − 14.4 year two-sided p value = 0.0001

therefore load = −110 + 0.681q + residual

= −110 + 0.681q + 28,250 − 14.4 year

= 28,140 − 14.4 year + 0.681q

12.2 Winter: P = 16, M = 34, S = -18
1 tie of 3, 2 ties of 2
Var[S] = 159.33
Z = −1.347
p = 0.18 very little evidence of downtrend in winter lead

Spring: P = 27, M = 38, S = −11
3 ties of 2, 1 tie of 5
Var[S] = 249
Z = 0.633
p = 0.53 no evidence of downtrend in spring lead

Summer: P= 16, M = 33, S = −17
1 tie of 4
Var[S] = 156.33
Z = −1.28
p = 0.20 very little evidence of downtrend in summer lead

Fall: P = 11, M = 37, S = −26
1 tie of 4, 1 tie of 2
Var[S] = 155.33
Z = 2.005
p = 0.045 fairly strong evidence of downtrend in fall lead

Seasonal Kendall: S = −72
VAR[S] = 720
Z = −2.646
p (2-sided)= 0.008

Thus, even though the evidence from no individual season was highly conclusive, the data from all seasons taken together provides highly conclusive evidence of a downtrend in lead.

12.3 Maumee River Trends in Total Phosphorus

12.3.1 Parametric analysis first: LOAD vs TIME
Simple linear regression: LOAD = 444 − 0.221 TIME
t = -0.42 p = 0.673
s = 20.59 R-sq = 0.1% R-sq(adj) = 0.0%

A boxplot of the residuals shows them to be terribly skewed. A transformation is required. Try logarithms. Then the regression equation is:

ln(LOAD) = 117 – 0.0592 TIME

t = –1.32 p = 0.189

s = 1.770 R-sq = 1.3% R-sq(adj) = 0.6%

There is a fairly normal distribution of residuals, so a test based on regression seems legitimate. Very weak evidence of trend -- (two-sided) p-value of 0.189. But are there strong flow and/or seasonal effects? A plot of the residuals versus log of streamflow (LQ) shows a strong dependence on flow. Removing this should greatly enhance the power to detect any trend which is present.

```
         -                                                   *   *
      2.4+                                                      *   *
         -
 RESID   -                                                **   *
         -                                                *
         -                                            2*  2
      1.2+                                           26 ***
         -                                       * *2*
         -                                 * * 22 2 ***
         -                                 * ** * 2*
         -                        **   *2* 2 *
      0.0+                          ***3***2
         -                     * *   *2* **    2
         -                   *  33*2 *2*
         -              * * ** 2   2     *
         -                2 43*      *
     -1.2+        **2**4**  *
         -       *2  2 **      *
         -         *
           --------+---------+---------+---------+---------+--------
lnQ
                  -1.2       0.0       1.2       2.4       3.6
```

Boxplots of residuals by month also show a strong seasonal cycle, high in the winter & spring, low in summer. The best model we could find includes time, ln(Q), $\ln(Q)^2$, and sine and cosine of $2\pi T$:

LLOAD = 83.3 – 0.0425 TIME + 1.08 lnQ + 0.0679 $\ln(Q)^2$ – 0.0519 SIN + 0.141 COS

Predictor	Coef	Stdev	t-ratio	p
Constant	83.32	22.06	3.78	0.000
TIME	-0.04250	0.01115	-3.81	0.000
lnQ	1.08175	0.04947	21.87	0.000
$\ln(Q)^2$	0.06789	0.01868	3.63	0.000
SIN	-0.05190	0.06252	-0.83	0.408
COS	0.14058	0.05441	2.58	0.011

s = 0.4398 R-sq = 94.1% R-sq(adj) = 93.9%

This is interpreted as a strong evidence of downtrend, with a p-value <.001 The slope (in log units) = –0.0425 per year. All coefficients are significant at $\alpha = 0.05$ except for sine($2\pi T$). The sine must either be left in, or both it and the cosine taken out. To test whether together they are significant, an F test is performed. The model without these terms, with the standard error s = 0.449, is:

LLOAD = 85.1 – 0.0434 TIME + 1.06 LQ + 0.0748 LQSQ

Predictor	Coef	Stdev	t-ratio	p
Constant	85.09	22.51	3.78	0.000
TIME	-0.04340	0.01138	-3.81	0.000
lnQ	1.05921	0.04518	23.44	0.000
$\ln(Q)^2$	0.07481	0.01865	4.01	0.000

s = 0.4490 R-sq = 93.7% R-sq(adj) = 93.6%

So the F test to compare these two models is:

$$F = \frac{(26.01 - 24.57) / 2.0}{0.193} = 3.73$$

Comparing to an F distribution with 2 and 127 degrees of freedom, the two-sided p-value is 0.027. Therefore reject the simpler model in favor of including the seasonal sine and cosine terms.

To predict estimates of load for the two times and two flow conditions above, natural logs of these values are input to the regression equation. The third column below reports the predicted logs of Load from the regression equation.

lnQ	Time	Predicted lnL	Bias-Corrected L
2.4	1972.5	2.3356	11.3852
0.0	1972.5	-0.6516	0.574136
2.4	1986.5	1.7406	6.27963
0.0	1986.5	-1.2467	0.316640

These predicitons must be transformed and corrected for bias. Using the Ferguson (MLE) bias correction, $0.5 \bullet s^2 = 0.5 \bullet (0.4398)^2 = 0.097$. So the bias correction equals exp(0.097), or about 10%. The four predicted total phosphorus loads are given above in the fourth column.

Therefore the percent change at high flow over the 14-year time period is:

(6.2763 – 11.3852) / 11.3852 = –0.448732

The change in percent per year is

–0.448732 • 100 / 14 = –3.205 . That is a –3.2% change in total P per year.

The same analysis at lower flow over the 14 years is:

(0.31664 – 0.574136) / 0.574136 = –0.4485, the same amount as at high flow.

Re-expressing the slope estimate in original units as a percent change, the average change equals –4.2% per year:

100 • [exp (–0.0425) – 1.0] = –4.16096

12.3.2 The nonparametric approach

The seasonal-Kendall test on the original observations, using 12 seasons (months): Tau = –0.06 with a p-value of 0.3835.

log(Load) = 0.505 – 0.046 •time ,

where time = 0 at the begining of the first year of the record (typically a water year), and time is in units of years.

Residuals from a regression of log(Load) versus logQ and $\log Q^2$ removes the effect of flow:

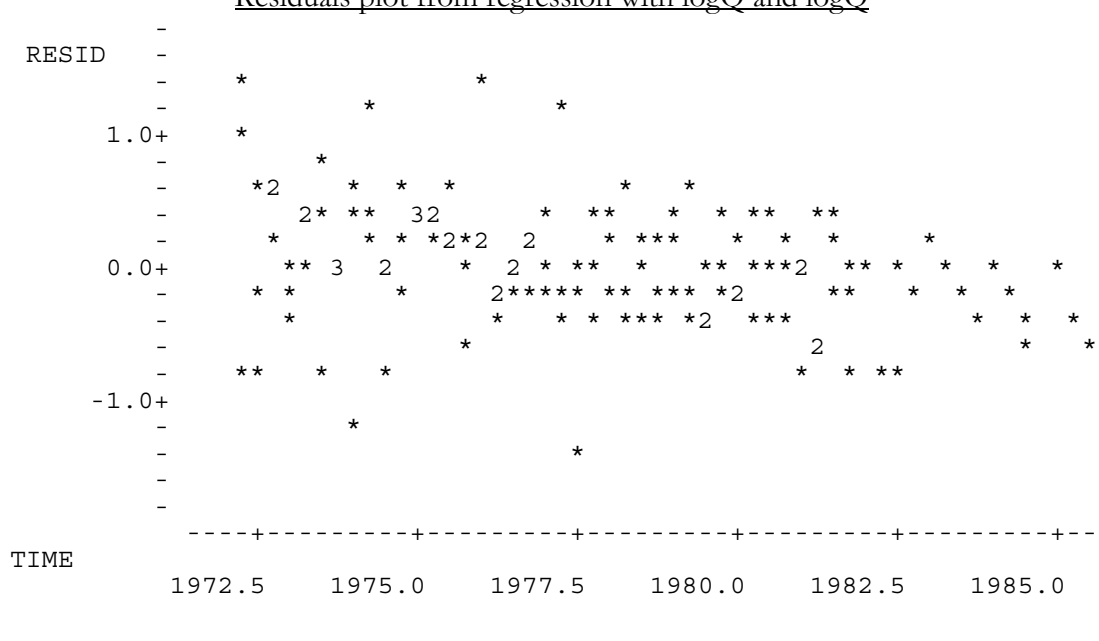
$\log(\text{Load}) = -0.745 + 1.06 \log Q + 0.0758 \log Q^2$

The S-K test on the regression residuals: tau = –0.25 with p = 0.0002

and log(Load) = 0.312 – 0.048 •time

So, if flow is not first removed, the significant trend would be missed. Both the Seasonal Kendall on the residuals and multiple regression give a highly significant p-value. The S-K slope is 4.8% rather than 4.16% because of the effect of the low residuals during 1972-1977.

<u>Residuals plot from regression with logQ and $\log Q^2$</u>



12.4 Not answered.

12.5 Nat answered.

Chapter 13

13.1 Not answered

13.2 Because there is only one reporting limit, Kendall's tau can easily be computed for this data: tau = −0.40 with p = 0.023. There is a significant decrease in TPT concentrations with depth.

13.3 Estimates of the four descriptive statistics for each of 5 multiple-threshold methods (see Helsel and Cohn, 1988) are:

Method	MEAN	ST.DEV.	MEDIAN	IQR
ZE (substitute zero)	12.36	75.48	0.00	1.10
HA (substitute 1/2 dl)	13.91	75.28	1.10	3.30
DL (substitute the dl)	15.45	75.19	1.30	4.10
MR (prob plot regression)	**12.57**	**75.44**	0.29	1.54
MM (lognormal MLE)	8.30	61.52	**0.34**	**1.62**

Because of the outlier at 560 μg/L the data have more skewness than a lognormal distribution, and methods which assume a lognormal distribution for all the data (MM) would not be expected to estimate moment statistics well. It is not surprising therefore that the MLE method produces moment estimates dissimilar to the other methods. We generally select the MR moment estimates and the MM quantile estimates (those printed in bold), due to the results of Helsel and Cohn (1988).

Chapter 14

14.1 a) Contingency table

Expected values E_{ij}

Trend in $C1^-$ (1974-81, α=0.1)

Δ road salt appl.	Down	No trend	Up	Totals
Down	5.44	23.84	16.71	46
No change	9.82	43.02	30.15	83
Up	13.73	60.13	42.14	116
Totals	29	127	89	**245**

Table of $\frac{(0-E)^2}{E}$

Trend in C1⁻ (1974-81, α=0.1)

Δ road salt appl.	Down	No trend	Up	Totals
Down	0.04	2.79	3.56	
No change	1.78	0.02	0.88	
Up	1.01	1.39	3.93	

$\chi^2 = 15.39$ df = 4 p = 0.004

The results indicate that the category of chloride trends is dependent on the category of salt applications, with a p-value of 0.004. Where increases in road salt occurred, there are more up trends and fewer down trends than would be expected from the marginal distributions of up trends and down trends. Where decreases in road salt occurred, there are fewer up trends than would be expected.

b) Kendall's tau

P = no. pluses = 5(44+25+51+55) + 32(25+55) + 14(51+55) + 44(55) = 7339

M = no. minuses = 32(14+10) + 9(14+44+10+51) + 44(10) + 25(10+51) = 3804

S = 7339 - 3804 = 3535.

$$\tau_b = \frac{3535}{\frac{\sqrt{(245^2 - (46^2 + 83^2 + 116^2))\,(245^2 - (29^2 + 127^2 + 89^2))}}{2}}$$

$$= 3535 / 18164 = 0.19$$

To test for significance,

$$\sigma_S \cong \sqrt{\frac{1}{9} * \left(1 - \left(.19^3 + .34^3 + .47^3\right)\right) * \left(1 - \left(.12^3 + .52^3 + .36^3\right)\right)}$$

$$= \sqrt{\frac{(0.85 * 0.81 * 245^3)}{9}} = 1061$$

and so $Z_S = 3534/ 1061 = 3.33$ and two-sided p= 0.0008.

The two variables are significantly and positively related.

c) Kendall's tau is more appropriate because

1. It includes the information that the variables are ordinal into the test. The p-value for Kendall's tau is lower than that for the contingency table, reflecting this additional information.
2. It provides a measure of the direction of association τ_b. Since τ is positive, the trends in Cl^- increase with increasing trends in road salt application.

14.2 Based on the tables below, there is a significant association between location of the well and the probability of detecting volatiles. The more protected downdip wells indeed have less chance of being contaminated than do the outcrop wells.

Expected values E_{ij}

Location	Non-detects	Detect VOC	Totals
Downdip	91.92	23.08	115
Outcrop	143.08	35.92	179
Totals	235	59	**294**

Table of $\frac{(0-E)^2}{E}$

Location	Non-detects	Detect VOC
Downdip	2.16	8.59
Outcrop	1.39	5.52

$\chi^2 = 17.647$ df = 1 p = <0.0001

14.3 To test for association, Kendall's tau-b is used because both time and concentration variables are ordinal. Computations are shown in the boxes below.

TBT Concentration

Year	≤ 200	> 200	A_i	a_i
1988	2	7	9	0.18
1989	9	13	22	0.43
1990	10	10	20	0.39
C_j	21	30	**51**	
c_j	0.41	0.59		**1.0**

The number of pluses $P = 2(13+10) + 9(10) = 136$
The number of minuses $M = 7(9+10) + 13(10) = 263$
So $S = 136 - 263 = -127$.

To compute the denominator of τ_b, $SS_a = 9^2 + 22^2 + 20^2 = 965.$
$SS_c = 21^2 + 30^2 = 1341.$

and $$\tau_b = \frac{-127}{\sqrt{\frac{(51^2 - 965)\,(51^2 - 1341)}{2}}} = \frac{-127}{717.9} = -0.18.$$

From equation 14.9 the approximate value of σ_S is

$$\sigma_S \cong \sqrt{\frac{1}{9} \bullet \left(1-(0.18^3+0.43^3+0.39^3)\right) \bullet \left(1-(0.41^3+0.59^3)\right) \bullet 51^3}$$

$$\cong \sqrt{\frac{(0.86)\bullet(0.73)\bullet 51^3}{9}} = \sqrt{9253} = 96.2$$

$$Z_S \cong \frac{-127+1}{96.2} = -1.31$$

and from a table of the normal distribution the one-sided p-value is $p = 0.095$. Therefore H_O: $\tau_b = 0$ is not rejected at α=0.05, but is for $\alpha = 0.10$. Thus there is weak evidence of a downtrend in TBT concentrations based on a split at 200 ng/L. Stronger evidence could be obtained by collecting data for subsequent years, or by obtaining better resolution of the data (the original data reported concentration values rather than a split at 200 ng/L).

Chapter 15

15.1 Logistic regression for the full model with four explanatory variables gives:

Variable Name	Parameter Estimate	Standard Error	Wald's t	p-value
Constant	−13.20539	3.55770	−3.71	0.0002
Thick	0.51527	0.15093	3.41	0.0004
Yields	0.42909	0.27484	1.56	0.0607
GW Qual	0.03035	0.32460	0.09	0.4642
Hazard	1.08952	0.29860	3.65	0.0002

with a likelihood ratio $lr_O = 49.70$ and $p<0.000$ as compared to the intercept-only model. However, two of the variables (Yields and GW Qual) have insignificant t-statistics. In the following model they are dropped, and lr_O recomputed:

Variable Name	Parameter Estimate	Standard Error	Wald's t	p-value
Constant	−10.89039	2.43434	−4.47	<0.0001
Thick	0.46358	0.13575	3.41	0.0004
Hazard	1.07401	0.28301	3.80	0.0001

with a likelihood ratio $lr_O = 52.54$ and $p<0.000$ as compared to the intercept-only model. The partial likelihood ratio to test whether the first model is significantly better than the simpler second model is:

$$lr = lr_{O(simple)} - lr_{O(complex)} = 52.54 - 49.70 = 2.84$$

which for a chi-square distribution with 2 degrees of freedom gives:

$$p = 0.242.$$

Therefore the two additional variables (Yields and GW Qual) do not appreciably add to the explanatory power of the model.

Index

3-dimensional rotation, 60
7-day 10-year low flow, 83
additive relation, 42, 142
    checking for, 186
adjacent values, 26
adjusted R^2, 399
adjusted variable Kendall test, 336
adjusted variable plots, 301
AIC, 399
Akaike's information criteria, 399
aligned-ranks, 192
alpha level, 107
alpha level, 68
alternate hypothesis, 104
analysis of variance, 158
    ANOVA table, 168
    assumptions of, 166
    multi-factor, 170
    on the ranks, 163
    on within-block ranks, 191
    one-factor, 164
    two-factor ANOVA table, 174
    two-way rank tests, 170
    unbalanced design, 179
angles, judgment of on graphs, 417
ANOVA. see analysis of variance
area, judgment of on graphs, 416
assumption of independence, 252
asymmetric confidence intervals, 69
    for the mean, 76
attained significance level, 108
autocorrelation. see serial correlation
bar chart, 62, 206
    grouped, 422
    stacked, 421
bias correction, 258
bisquare weight function, 284
block effect, 196
blocking, 181, 197
Blom plotting position, 24
boxplots, 24, 39, 423
    side-by-side, 129
boxplots, 128, 207
bulging rule, 230
business graphics, 412
categorical variables
    Kruskal-Wallis test for, 382
categories, 19
censored data, 3, 124, 128
    guidelines for use, 373
    in trend tests, 353
    nonparametric tests, 366
    parametric tests, 366
    regression, 370
characteristics of data, 2
chemometrics, 372
coefficient of determination, 228
coefficient of skewness, 11
Cohen's method, 353, 360
color, use of in graphs, 412
comparing among distributions, 35
comparison value, 186
compliance with water quality standards, 83
confidence intervals, 66
    for percentiles, 82
    for percentiles (nonparametric), 84

confidence intervals**, cont**.
for percentiles (parametric), 90
for regression line, 242
for skewed data, 69
for the mean, 75
for the median, 70
confidence level, 68
constant variance assumption, 255
contingency tables, 378
for censored data, 372
continuity correction, 122, 141, 147
control chart, 93
Cook's D, 250
correction for ties
Kendall's tau;, 215
rank-sum test, 123
correlation, 209
monotonic, 210
correlation coefficient, 210
linear, 218
nonparametric, 212
critical values
lower, 113
upper, 113
cumulative distribution function, 23
cumulative line graph, 419
cumulative logits, 406
Cunnane plotting position, 22, 24, 114
degrees of freedom, 160, 167, 297, 310, 380
detection limit, 3, 128
guidelines for use of data, 373
more than one, 354
trend tests for data below, 353
DFFITS, 250
difference between group means, 135
confidence interval for, 135
differences between groups
estimators, 132
discriminant function analysis, 402
dot and line plot, 36, 38
dot charts, 423
Duncan's multiple range test, 199, 200
Durbin-Watson statistic, 253
efficiency
asymptotic relative, 102
Kendall vs. OLS, 268
equal variance, 124
error rate, 107, 110
overall, 199
pairwise, 199
error sum of squares, 167, 196
exogenous variables in trend tests, 330
expected value, 224
explanatory variable, 99, 158, 222
extension of records, 278
factor, 158
factorial ANOVA, 170
assumptions of, 173
far-out values, 26
fixed effects, 180
flood-frequency, 24
flow-duration, 24
framed rectangle, 421
Friedman test, 170, 187, 192
large sample approximation, 188
F-test, 168, 174
geometric mean, 7, 73
estimation for censored data, 361
geometric mean functional regression, 276
graphical analysis, 19
graphical comparisons, 35
graphical methods, 128
graphics, 205
graphs
pie chart, 416
angle and slope, 417
boxplots, 423
cumulative line graphs, 419
dot charts, 423
framed rectangles, 421

graphs, **cont.**
grouped bar charts, 422
hidden scale breaks, 428
judgments of length, 420
misleading, 424
overlapping histograms, 429
precision of, 411
stacked bar charts, 422
use of color, 412
use of numbers on, 427
use of perspective, 424
use of shading, 413
Gringorten plotting position, 24
Gumbel distribution, 90
harmonic mean, 199
Hazen plotting position, 24
heavy tails, 2
heteroscedasticity, 124, 230, 255, 280
hinges, 25
histograms, 206, 429
histograms, 19, 36
Hodges-Lehmann estimator, 132, 155
confidence interval for, 133, 155
for step trend, 349
homoscedasticity, 51
homoscedasticity, 13
hypothesis tests, 97, 109
choice of, 101
classification, 99
classification of, 106
exact, 103
with censored data, 365
independent groups, 117
inequality of variance, 124
influence, 4, 248, 250
interaction, 172
intercept, 226
confidence interval for, 240
deletion of, 240
nonparametric, 268
interquartile range, 9, 24
interval estimates, 66
invariance to rotation, 280
IQR, 9
iteratively weighted least squares, 283
joint probability, 379
Kendall's nonparametric line;, 370
Kendall's nonparametric line; .i.Sen slope estimate; .regression
nonparametric;, 266
Kendall's S statistic, 272
Kendall's tau, 326
for categorical variables., 386
for censored data, 372
large sample approximation;, 213
tie correction;, 215
Kendall's tau; .i.tau;, 212
Kendall's tau-b;)., 386
kite diagram, 54
Kruskal-Wallis test, 158, 159
for categorical variables, 382
large sample approximation, 160
rank transform approximation, 163
ladder of powers, 14, 15, 31, 119, 230
large sample approximation, 103, 121
least normal squares, 279
least significant range, 198
least squares, 228
length, judgment of on graphs, 420
leverage, 248
likelihood ratio, 397
likelihood-R^2, 398
line of organic correlation, 276
linear regression. see regression
linearity, 13
linearity, 46
LOC, 276
log likelihood, 397

logistic regression, 395
for censored data, 371
for multiple responses, 403
logistic transformation, 396
logit, 396
loglinear models, 390
lognormal, 2
lognormal distribution, 73
LOWESS, 47, 288
use in trend tests, 335
LOWESS, 325
MAD, 10, 284
Mann-Kendall trend test, 326
with censored data, 354
Mann-Whitney test, 118
MARA, 192
marginal probability, 379
mass transport, 258
matched-pair tests
graphical presentation of, 152
maximum-likelihood, 360
tobit regression, 370
mean
asymmetric confidence interval for, 76
confidence interval for, 75
mean difference, 157
mean square error, 228
mean squares, 167, 174
measures of location, 3
median, 6
confidence interval for, 70
test for differences in, 118, 159
Median Aligned-Ranks ANOVA, 192
median difference, 154
median polish, 182
mixed effects, 181
MLE, 360
tobit regression, 370
mode, 7
monotonic correlation, 210
monotonic trend, 327
MOVE, 276
moving average, 286
moving medians, 286
multiple comparison test
graphical display of, 208
multiple comparison tests, 197
for categorical variables, 385
nonparametric, 198, 203
multiple regression, 237
use as a trend test, 337
multiplicative relation, 43
multiply-censored data, 353
hypothesis tests for, 368
nonparametric tests for, 368
multivariate graphical methods, 52
nominal response variables, 406
non-normality, 92, 124
nonparametric interval estimate, 70
nonparametric prediction intervals, 77
nonparametric test for percentiles, 86
nonparametric tests, 101
comparison to parametric, 102
for censored data, 366
normal distribution, 2, 26, 31, 113
normal probability plot, 24, 114
normal quantiles, 28, 114
normal scores, 114
normality
of test statistics, 121
test of, 102, 113, 166
normality assumption, 150
null hypothesis, 104, 107, 108
not rejecting, 109
rejecting, 109
odds ratio, 396
OLS, 222
one-sided p-value, 112
one-sided tests, 105, 109
ordinal variables, 386

ordinary least squares, 222
outliers, 12, 92
    tests for, 92
outliers, 2, 31, 38, 127, 250
outside values, 26
overall error rate, 199
paired observations, 138
paired t-test, 148
    assumptions of, 148
    computation, 149
pairwise comparisons, 199
pairwise error rate, 199
parametric c. i. for the median, 73
parametric prediction intervals, 80
parametric tests, 100
    comparison with nonparametric, 102
    with censored data, 365, 366
partial likelihood ratio, 398
Pearson Type III distribution,, 90
Pearson's r, 218
percent exceedance, 30
percentiles
    confidence interval for, 83
    parametric tests for, 91
    water quality, 83
percentiles, 9
perceptual tasks for interpreting graphs, 412
perspective, use of in graphs, 424
pie chart, 62
pie charts, 416
Piper diagram, 58, 59
Piper smooth, 59
plotting position, 22, 23, 30, 114
point estimates, 66
polar smooth, 48, 292
population, 2
positive skewness, 10
power, 95, 100, 107
    lack of, 124, 127
    loss of, 102, 150
power, 102
power transformation
    effect on paired t-test, 151
    effect on rank-sum test, 118
    effect on signed-rank test, 144
power transformations, 177
    avoiding, 265
    effect on t-test, 128
    invariance to, 327
    WLS as an alternative, 282
PPCC test, 113
precision of graphs, 411
prediction interval, 66, 243
    asymmetric, 81
    nonparametric, 77, 79, 81, 244
    one-sided, 79
    parametric, 80, 81
    symmetric, 81
    two-sided, 78
prediction residual, 249
PRESS statistic, 249
principal components analysis, 59
probability paper, 29
probability plot, 27, 31, 35, 41
probability plot correlation coefficient, 35
probability plot correlation coefficient, 113
profile plot, 53
p-value, 108, 111
    one-sided, 113
    two-sided, 112, 113
Q-Q plot, 42, 43, 128, 129
    construction, 45
quality control, 93
quantile plot, 22, 206
quantile-quantile plot, 129
quantiles, 22, 83
quantiles, 42
quartile, 9
quartile skew coefficient, 11
r squared, 228

random effects, 180
randomized complete block design, 182, 187
rank transform test, 194
rank transformation test, 123, 170, 203
ranks, 7
ranks, 104
rank-sum test, 110, 118
    an alternative to logistic regression, 402
    as a test for trend, 349
rating curve, 258
record extension, 278
regression, 221
    for censored data, 369
regression
    as a test for trend, 328
    assumptions, 224
    confidence interval on mean response, 242
    diagnosing problems, 232
    guide to model selection, 263, 316
    hypothesis testing, 238
    non-normal residuals, 268
    nonparametric, 266
    normality assumption, 236
    robust, 269, 283
    validation of equation, 249
regression diagnostics, 238, 246
rejection region, 113
reliability, 66
replicates, 194
replication
    ANOVA without, 194
residuals, 31, 226
    prediction, 249
    standardized, 249
    studentized, 249
    testing for normality, 236
    use in trend tests, 332
residuals plot, 187, 232
    trend, 234
resistant, 6
response variable, 99, 222
rho, 217
risk tolerance, 108
RMSE, 358
robust, 11
robust regression, 269
root mean squared error, 358
sample, 2
sample size, estimating, 95
sampling design, 95
Satterthwaite's approximation, 126
scale breaks, 428
scatterplot matrix, 61
scatterplots, 46, 423
schematic plot, 26
seasonal Kendall test, 339
seasonal rank-sum test, 350
seasonal variation, 234
    differences among seasons, 345
    graphics for display of, 344
    modeling, 338
    use of periodic functions, 342
Sen slope estimate, 266
serial correlation, 252
    remedies, 254
shading on graphs, 413
Shapiro-Wilk test, 115
sign test, 138, 187
    computation, 138
    large sample approximation, 141
signed-rank test, 142, 192
    large sample approximation, 146
    rank transform approximation, 148
significance level, 107, 110
simple boxplot, 25
skew, 10
skewed data, 124
skewness, 2, 31, 69
skewness, 38, 40, 127

slope, 226
confidence interval for, 240
judgment of on graphs, 417
test of significance, 238
smearing estimator, 259
smooth, 46, 47, 48
lower, 292
LOWESS, 286
middle, 286
outer, 293
upper, 292
smooth, 47
smoothing, 286
smoothness factor, 289
spatial trend, 326
Spearman's rho, 217
spread, 8, 51
stacked bar chart, 62
standard deviation, 9, 36
standard error, 36, 228
standardized residual, 249
star diagram, 54, 58
statistical maps, 413
statistical tables, 113
stem and leaf diagram, 20
step trend, 349
when to test for, 351
Stiff diagram, 53, 58
studentized range, 198
student's t statistic, 75
sum of squares, 196
summary statistics, 10
for censored data, 358
with multiple reporting limits, 364
sums of squares, 173
symmetric confidence intervals, 68, 75
symmetry, 13
assumption of, 151
table of test statistic quantiles, 110
tables, deficiencies of, 410
tails of the distribution, 10, 31
target population, 2
tau, 212
t-distribution
noncentral, 90
test statistic, 108
Theil slope estimate
computation, 266
confidence interval for, 273
efficiency, 268
for trends, 330
with censored data, 354
tie correction
for tests with censored data, 354
tobit regression, 366, 370
tolerance intervals, 83
tolerance probability, 96
transformation bias
in regression, 258
of MLE, 360
transformations, 31, 103, 166, 177, 255
consequences in regression, 256
transformations, 13, 230
t-ratio, 232
trend
exponential change, 347
trend analysis
including exogenous variables, 330
nonparametric, 335
step trends, 349
use of transformations, 347
with censored data, 353
trend slope, 330
seasonal, 341
trilinear diagram, 57
trimmed mean, 7
truncated boxplot, 26

t-test, 103, 124
as a test for trend, 349
assumptions of, 124
computation of, 125
for multiple comparisons, 199
on ranks, 127
problems with, 124
violation of assumptions, 127
Tukey's multiple comparison test, 199, 201
two-factor ANOVA, 193, 194
two-sided tests, 105, 111
Type I error, 107
Type II error, 107
unequal sample sizes, 179
unequal variances, 126
variance, 9
confidence interval for, 240
violation of test assumptions, 150
Wald's t-statistic;, 398
Weibull plotting position, 24
weight function, 288
weighted least squares, 281
whisker, 25
WLS, 281